UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga

Uporaba razpoloženja objav na omrežju Twitter za napovedovanje vrednosti delnice TSLA

(Use of sentiment from Twitter posts for predicting the TSLA share value)

Ime in priimek: Sebastijan Peruško
Študijski program: Računalništvo in informatika
Mentor: izr. prof. dr. Marko Tkalčič
Somentor: doc. dr. Branko Kavšek

Koper, september 2022

Ključna dokumentacijska informacija

Ime in PRIIMEK: Sebastijan PERUŠKO

Naslov zaključne naloge: Uporaba razpoloženja objav na omrežju Twitter za napovedovanje vrednosti delnice TSLA

Kraj: Koper

Leto: 2022

Število listov: 66 Število slik: 19 Število tabel: 10

Število referenc: 38

Mentor: izr. prof. dr. Marko Tkalčič

Somentor: doc. dr. Branko Kavšek

Ključne besede: računalništvo, strojno učenje, časovne vrste, delnice, napovedovanje

Math. Subj. Class. (2020):

Izvleček:
V zaključnem delu se osredotočamo na uporabo razpoloženja objav na omrežju Twitter za napovedovanje zadnjega tečaja delnice TSLA. V koraku analize razpoloženja smo analizirali in pridobili polarnost razpoloženja 1.234.137 tvitov v obdobju med 1. decembrom 2019 in 1. decembrom 2021. Dobljene vrednosti razpoloženja smo združili in jih nato uporabili za izračun Pearsonovega korelacijskega koeficienta, pri čemer smo dobili 0,678 linearne korelacije med razpoloženjem v pridobljenih tvitih in vrednostjo zadnjega tečaja delnice. Pri napovedovanju vrednosti delnice smo analizirali lastnosti časovnih vrst, v katere spadajo trend, sezonskost in cikli. Poleg tega pa smo uporabili in primerjali različne algoritme in uporabili hibridne modele strojnega učenja. Pri napovedovanju vrednosti delnice smo ugotovili, da je napovedovalni model vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji le na preteklih vrednostih delnice. Pri uporabi razpoloženja smo uspeli napovedati vrednosti delnice z R^2 = 0,985. Nazadnje bomo predstavili tudi napovedovanje gibanja vrednosti delnice, pri čemer smo ugotovili, da je napovedovalni model gibanja vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji le na preteklih vrednostih delnice. Pri uporabi razpoloženja je točnost napovedovanja gibanja vrednosti zadnjega tečaja delnice 67,99-odstotna.

Key words documentation

Name and SURNAME: Sebastijan PERUŠKO

Title of final project paper: Use of sentiment from Twitter posts for predicting the TSLA share value

Place: KOPER

Year: 2022

Number of pages: 66 Number of figures: 19 Number of tables: 10

Number of references: 38

Mentor: Assoc. Prof. Marko Tkalčič, PhD

Co-Mentor: Assist. prof. Branko Kavšek, PhD

Keywords: computer science, machine learning, time series, shares, forecasting

Math. Subj. Class. (2020):

Abstract: In this final work, we focus on using the sentiment from Twitter posts for predicting the value of TSLA stock. In the sentiment analysis step, we analyzed and extracted the sentiment polarity of 1,234,137 tweets during the period between December 1, 2019 and December 1, 2021. We aggregated the obtained sentiment values, and then we used them to calculate the Pearson correlation, obtaining 0.678 linear correlation between the sentiment in the extracted tweets and the stock closing value. In predicting the share value, we analysed the properties of the time series, which include trend, seasonality and cycles. In addition, we did apply and compare different algorithms and used hybrid machine learning models. When predicting the value of a share, we found that a predictive model of the value of a share that incorporates sentiment outperforms a model based only on the past values of the share. When using sentiment, we were able to predict the value of a share with $R^2 = 0.985$. Finally, we also present the prediction of the share closing value movement, where we find that the predictive model of the share closing value movement incorporating sentiment outperforms the model based only on the past values of the share. When sentiment is used, the accuracy of predicting the share price movement is 67.99%.

Zahvala

Zahvaljujem se mentorju izr. prof. dr. Marku Tkalčiču za oblikovanje izvirne ideje in usmerjanje do končne rešitve te zaključne naloge. Zahvaljujem se tudi somentorju doc. dr. Branku Kavšku za pomoč in podane smernice, ki so me vodile pri izdelavi drugega dela zaključne naloge.

Zahvaljujem se tudi mami Orjani in sestri Veroniki za potrpežljivost in pomoč v vseh teh letih.

Kazalo vsebine

Kazalo tabel

Kazalo slik

Seznam uporabljenih kratic

Kratica	Angleško	Slovensko
CSV	Comma-separated values	Vrednosti, ločene z vejico
HTML	HyperText Markup Language	Jezik za označevanje nadbesedila
MSE	Mean Squared Error	Povprečna kvadratna napaka
NLTK	Natural Language Toolkit	Orodje za naravni jezik
RMSE	Root Mean Squared Error	Koren povprečne kvadratne napake
SVM	Support vector machine	Metoda podpornih vektorjev
SVR	Support vector regression	Regresijska metoda podpornih vektorjev
VADER	Valence Aware Dictionary and sEntiment Reasoner	Valenčno podprt slovar in utemeljitve čustev
Zadnji tečaj	Closing price	Cena delnice na koncu trgovalne seje

1 Uvod

Danes na borzi kotirajo številna podjetja, in sicer od najmanjših do velikanskih multinacionalk. Vsak dan se na borzi investira in prenese na milijarde evrov. Poleg tega je le-ta predmet novic, ki jih slišimo vsak dan in je sestavni del gospodarstva. Napovedovanje vrednosti delnic je že od nastanka borznega trga ena od najbolj zanimivih tem in izzivov, ki privablja delničarje in raziskovalce. Sprva je veljalo, da se vrednost delnic spreminja na podlagi naključne hoje in trdi, da je sprememba cene med katerimakoli dvema zaporednima obdobjema naključna spremenljivka, ki je neodvisna od predhodnih sprememb cen [15]. Nadaljnje raziskave na področju napovedovanja delnic so temeljile na hipotezi učinkovitega trga. Namreč, vrednosti delnic bi morale slediti vzorcu naključnega gibanja, torej spremembe cen so zgolj posledica novih informacij in so neodvisne od obstoječih informaciji in zato ne bi smeli napovedati z več kot 50-odstotno natančnostjo [24]. Sedanje raziskave pa zajemajo veliko širše področje, imenovano vedenjska ekonomija, ki temelji na hipotezi, da so finančne odločitve v veliki meri odvisne od čustev in razpoloženja ljudi [7].

Z razvojem računalnikov in stojnega učenja je bilo tudi napovedovanje vrednosti delnic predmet obsežnih raziskav. Hkrati pa se vse bolj širijo družbena omrežja, ki vsakemu posamezniku in skupini omogočajo, da izrazi svoje stališče o določeni temi. Družbena omrežja so se izkazala kot vir številnih koristnih informacij, vključno z napovedovanjem določenih dogodkov, kamor spadajo tudi delnice. Eden od načinov analize sporočil na družbenih omrežjih je analiza razpoloženja, ali rudarjenje mnenj, ki analizira in ekstrapolira čustva osebe glede določene stvari.

V zadnjih letih so raziskave pokazale, da je mogoče napovedati vrednost delnic z analizo sentimenta objav v družbenih omrežij, kot sta Twitter in Reddit in celo, da je mogoče napovedati vrednost delnic z analizo sentimenta na podlagi naslovov novic. Pravzaprav so enega od najbolj znanih delov ustvarili Bollen J. in drugi v [7], ki so uspeli dokazati, da obstaja korelacija med borznim indeksom Dow Jones Industrial Average in sentimenti v tvitih.

Od tukaj tudi izvira ideja zaključne naloge, ki se bo osredotočala na raziskovanje področja vedenjske ekonomije za napoved gibanja in vrednosti delnice. Namreč, cilj v

zaključni nalogi je razumeti, ali je gibanje razpoloženja objav na omrežju Twitter povezano z gibanjem vrednosti delnice in na podlagi le-tega izboljšati napoved vrednosti in gibanja vrednosti delnice. V zaključni nalogi se bomo osredotočili le na napovedovanje zadnjega tečaja delnice, ki je vrednost delnice na koncu trgovalne seje oziroma cena ob zaprtju borze.

1.1 Struktura zaključne naloge

Zaključna naloga je razvrščena v več poglavji, v katerih je predstavljen potek dela na področju napovedovanja vrednosti delnic.

V 2. poglavju bomo predstavili sorodno delo na področju napovedovanja vrednosti delnic na podlagi vedenjske ekonomije. Poleg tega bomo demonstrirali tudi raziskovalna vprašanja zaključne naloge.

V 3. poglavju bomo predstavili teoretično ozadje in uporabljena orodja zaključne naloge. To poglavje obsega temeljne pojme ter temelje časovnih vrst in strojnega učenja. Poleg tega pa bodo predstavljena orodja in tehnologije, ki so bile uporabljene v tej zaključni nalogi.

V 4. poglavju bomo predstavili predštudijo, ki je bila narejena na področju napovedovanja vrednosti delnice. To poglavje zajema postopek pridobivanja, čiščenje in procesiranja podatkov o dogodkih, ki so se odvijali v Združenem kraljestvu. Študija je bila odpuščena za trenutno zaključno nalogo, zaradi ne najdene povezave med vrednostmi delniškega trga in razne vrste dogodkov.

V 5. poglavju bomo predstavili metodologijo glavne študije diplomske naloge. To poglavje zajema postopek pridobivanja, čiščenja in procesiranja podatkov o delnici in o tvitih. Nato bomo predstavili korake sentimentalne analize objav na omrežju Twitter.

V 6. poglavju bomo predstavili postopek določanja povezave med razpoloženjem objav na omrežju Twitter in gibanjem vrednosti delnice. To poglavje zajema iskanje optimalnega intervala in optimalne strategije združevanja razpoloženja objav na omrežju Twitter s pomočjo Pearsonovega korelacijskega koeficienta. Na koncu bodo predstavljeni tudi rezultati korelacije med vrednostjo delnice podjetja Tesla in razpoloženjem objav Twitterja.

V 7. poglavju bomo predstavili napovedovalni model, ki se osredotoča na napovedovanje vrednosti delnice. Vrednost delnice bomo namreč napovedali na podlagi značilnosti časovnih vrst in algoritmov strojnega učenja. V tem poglavju se bomo osredotočili na primerjavo med napovedovalnim modelom, ki temelji zgolj na podatkih, ki izhajajo iz vrednosti delnice in napovedovalnim modelom, ki temelji na razpoloženju objav na omrežju Twitter.

V 8. poglavju bomo predstavili napovedovalni model, ki napoveduje gibanje vredno-

sti delnice. Gibanje vrednost delnice bomo namreč napovedali na podlagi značilnosti časovnih vrst in algoritmov strojnega učenja. V tem poglavju se bomo osredotočili na primerjavo med napovedovalnim modelom, ki temelji zgolj na podatkih, ki izhajajo iz vrednosti delnice, in napovedovalnim modelom, ki temelji na razpoloženju objav na omrežju Twitter. Poleg tega bomo ta model testirali in simulirali nakup in prodajo delnice in na koncu predstavili hipotetičen dobiček.

V 9. poglavju bomo predstavili zaključek zaključne naloge, pri čemer bomo opisali dosežke trenutne zaključne naloge in njene omejitve. Poleg tega bomo prikazali možnosti za nadaljnje delo na tem področju.

2 Pregled literature

V tem poglavju se osredotočamo na sorodno delo na področju analize razpoloženja in uporabo le-tega za napovedovanje vrednosti delnice. Poleg tega bomo predstavili zasnovo in raziskovalna vprašanja zaključne naloge.

2.1 Napovedovanje vrednosti delnic s pomočjo analize razpoloženja objav s Twitterja

V raziskavi [21] avtorjev Pagolu V. S. in drugi so pretvorili napovedane delnice v klasifikacijski problem. Če je namreč cena delnice trenutnega dneva večja od prejšnjega dneva, bo dan označen s številčno vrednost 1, v nasprotnem primeru pa z 0. Poleg tega so bili tviti glede na razpoloženje razvrščeni na negativne, nevtralne in pozitivne, pri čemer so se uporabile N-gram in Word2vec tehnike za predstavitev besedila. Rezultati klasifikatorja pri napovedi vrednost delnice kažejo 69,01-odstotno točnost, pri čemer je bil uporabljen algoritem strojnega učenja logistična regresija z 80-odstotkov podatkov namenjenih za učno množico. Pri uporabi algoritma LibSVM je 90-odstotkov podatkov namenjenih za učno množico, točnost je 71,82-odstotna.

Ena najpomembnejših raziskav na tem področju je [7], avtorjev Bollen J. in drugi je pokazala, da obstaja povezava med Dow Jones Industrial Average in razpoloženjem na Twitterju. Upoštevani so bili samo tviti, ki izrecno izražajo občutke ljudi in bolj natančno tviti, ki vsebujejo kombinacijo ključnih besed "I am", "I feel", "I don't feel", "makes me"in tako naprej. Poleg tega so bili za izognitev spam objav odstranjeni tviti, ki ustrezajo regularnim izrazom "http:"ali "www.". Posledično so bili tviti analizirani s strani dveh orodji, ki sta OpinionFinder in Google-Profile of moods(GPOMS). Z uporabo Opinionfinder so bile posamezne besede tvitov klasificirane v negativne in pozitivne, nato pa je bilo izračunano razmerje med pozitivnimi in negativnimi besedami v tvitih. Za zajemanje celotnega spektra človeškega razpoloženja pa je bilo uporabljeno orodje Google-profile of moods, ki klasificira tvite na 6 stanj, ki so miren, pozoren, prepričan, vitalen, prijazen in srečen. Grangerjev test vzročnosti in napovedovalni model samoorganizirajoča se ohlapna nevronska mreža (ang. self-organized fuzzy neural network) sta bila testirana na podatkih delnice in atributov, ki izhajajo iz OpitionFinder

in Google-Profile of moods in so ugotovili, da uporaba določenih podatkov iz analize razpoloženja izboljša napoved. Rezultati so namreč pokazali, da napovedovalni model SOFNN s kombinacijo stanj GPOMS miren in srečen in prejšnje vrednosti delnice napovedujejo naraščanje in padanje vrednosti indeksa Dow Jones Industrial Average z 87,6-odstotno natančnostjo. Druga stanja razpoloženja OpitionFindera in Google-Profile of moods ne vplivajo na izboljšavo napovedi naraščanja, ali padanja indeksa.

V raziskavi [6] avtorjev Bing L. in drugi so klasificirali razpoloženje tvitov na skrajno negativne, negativne, nevtralne, pozitivne in skrajno pozitivne z uporabo leksikalne aplikacije SentiWordNet 3.0. Napoved pa je zajemala naraščanja, padanja, ali nespremenjenost vrednosti delnice 30 podjetji, ki kotirajo na borzi NASDAQ. Pokazali so, da je za nekatere delnice težje napovedati njihovo prihodnjo vrednost z analizo razpoloženja na Twitterju. Namreč IT, medijska in finančna podjetja so bolj občutljiva, medtem ko so farmacevtska, proizvodna in energetska podjetja manj občutljiva na razpoloženje javnosti na Twitterju.

V [4] so avtorji Skuza M. in drugi poleg napovedi vrednosti delnice APPL podjetja Apple primerjali vrste tvitov, ki na najboljši način napovedujejo vrednosti delnice. Uporabili so namreč dve zbirki podatkov, od katerih je ena vsebovala samo tvite, ki vsebujejo besedo APPL, druga pa tvite, ki vsebujejo ime podjetja "Apple". Rezultati so pokazali, da so napovedi, izvedene z modeli, usposobljenimi na podatkovnih množicah s sporočili, ki vsebujejo simbol delnice podjetja(npr. $AAPL) bolj natančne v primerjavi z modeli, usposobljenimi na podatkovnih množicah, ki vsebujejo celotno ime podjetja (npr. Apple).

V raziskavi [23] avtorjev Pyeong Kang Kim D. in drugi so analizirali povezavo Twitter objav avtorja generalnega direktoja Tesle Elona Muska in cene delnice TSLA. Tviti so bili klasificirani na negativno, nevtralno in pozitivno z uporabo pristopa strojnega učenja SVM. Rezultati so pokazali, da obstaja korelacija med objavami Elona Muska in vrednostjo delnice. Rast števil pozitivnih tvitov je povezana z rastjo cene delnice in nasprotno velja za negativne tvite.

Iz tovrstnih sorodnih del, ki so se osredotočala na vpliv razpoloženja na delniškemu trgu vidimo, da je razpoloženje uporabnikov povezano z gibanjem delnice in predvsem, da je mogoče uporabiti to razpoloženje za napovedovanje vrednosti delnice.

2.2 Opredelitev raziskovalnih vprašanj

Raziskava [23] opisana v prejšnjem razdelku opisuje vpliv tvitov glavnega direktorja podjetja Tesla. V temu zaključnemu delu se bomo pri razširitvi te tematike osredotočali na širše področje, pri čemer bomo uporabili tvite vseh uporabnikov, katerih objave so vsebovale ključne besede glede podjetja Tesla. Poleg tega pa bomo zgradili napovedovalni model na podlagi razpoloženja teh uporabnikov. V tem zaključnem delu se namreč osredotočamo na razvoj modela strojnega učenja za napovedovanje cene delnice naslednjega dneva s pomočjo analize razpoloženja. Izbrana delnica je TSLA podjetja Tesla, analiza razpoloženja pa bo izvedena na podlagi tvitov na družbenem omrežju Twitter.

Raziskovalna vprašanja (RV), na katera bomo odgovorili v tem zaključnem delu, so:

Prvo raziskovalno vprašanje je bilo:

- RV1: Ali je gibanje razpoloženja na družbenih omrežjih po nekem dogodku povezano z gibanjem vrednosti borznega indeksa države, v kateri se je dogodek zgodil?

V predštudiji, ki je opisana v poglavju 4, rezultati niso pokazali obstoja te povezave. Na podlagi teh dognanj smo zastavili naslednja raziskovalna vprašanja:

- RV2: Ali je gibanje razpoloženja na omrežju Twitter povezano z gibanjem vrednosti delnice TSLA?
- RV3: Ali je napovedni model vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji samo na preteklih vrednostih delnice?
- RV4: Ali je napovedni model gibanja vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji samo na preteklih vrednostih delnice?

3 Teoretično ozadje in uporabljena orodja

V tem poglavju predstavljamo temeljno tehnologijo, orodja in teorijo, ki bodo vključene v nadaljnje faze zaključne naloge. Predstavili bomo ključne pojme sentimentalne analize, časovnih vrst, strojnega učenja in ostalih uporabljenih tehnologij.

3.1 Twitter

Twitter [1] predstavlja enega od glavnih družbenih omrežji na svetu. Poleg velikega števila uporabnikov je za našo študijo in za tovrstne raziskave še posebej privlačen način ustvarjanja in objavljanja sporočil. Twitter namreč temelji na sporočilih, dolgih do 280 znakov, ki lahko vsebujejo slike ali videoposnetke. Ker je število znakov omejeno, morajo biti sporočila kratka, zato je lažje sklepati o uporabnikovem razpoloženju glede določene teme. Poleg tega je Twitter glavno družbeno omrežje, uporabljeno s strani generalnega direktorja Tesle Elona Muska, ki na svoje tvite pritegne veliko pozornosti ter posledično všečkov in odzivov.

3.1.1 Osnovna sintaksa Twitterja

V tem razseku bomo predstavili osnovno sintakso in pojme, ki so uporabljeni na družbenem omrežju Twitter, kar nam bo kasneje pomagalo pri predstavitvi nadaljnjih tem. V sliki 1 je prikazana ena objava na omrežju Twitter, pri čemer je osnovna sintaksa sledeča:

- Simbol #, imenovan hashtag in beseda, ki mu sledi brez presledka, se uporablja pri označevanju tem v objavi.
- Simbol @ in beseda, ki mu sledi brez presledka, se uporablja pri omenjanju določenega uporabnika ali uporabnikov v objavi.
- Simbol $, imenovan cashtag in beseda, ki mu sledi brez presledka, se uporablja pri označevanju delnic podjetji.

Slika 1: Prikaz ene objave na omrežju Twitter.

3.2 Analiza razpoloženja

Analiza razpoloženja je postopek analize določenega besedila za razumevanje razpoloženja pisatelja glede določene teme. Cilj analize razpoloženja je klasificirati besedilo na negativno, nevtralno in pozitivno, ali pa dodeliti določeno oceno, ki označuje polarnost besedila [36]. V tabeli 1 so prikazani primeri besedil in njihovo razpoloženje, ki je klasificirano na negativno, nevtralno in pozitivno.

Tabela 1: Primeri sentimentalne analize na določenih besedilih.

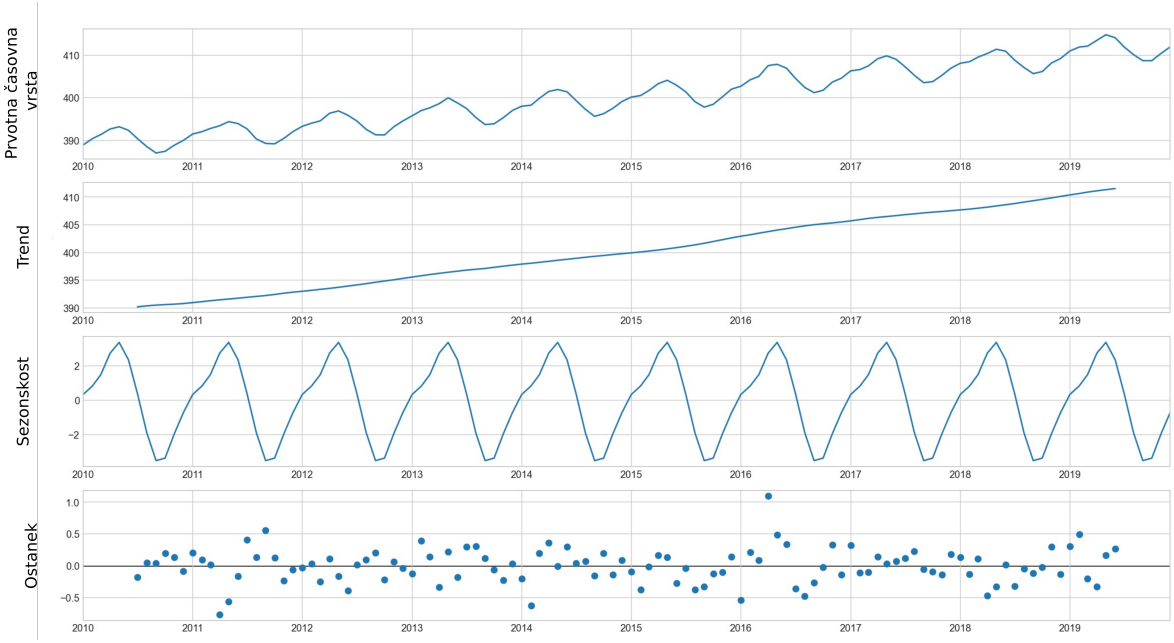
Besedilo	Razpoloženje
Tesla bo zagotovo propadla.	Negativno
Podjetje Tesla lahko postane pomembno, ali pa tudi ne.	Nevtralno
Tesla bo postala najpomembnejše avtomobilsko podjetje na svetu.	Pozitivno

3.3 Časovne vrste

Časovne vrste so množica podatkovnih točk, ki so razporejene v kronološkem vrstnem redu in prikazujejo določeni pojav [37]. Časovne vrste so del našega vsakdanjega življenja in so na primer temperature zraka za določene dni, ali pa število COVID-19 primerov v Sloveniji od začetka epidemije. Časovne vrste lahko razčlenimo na 4 glavne komponente, ki so trend, sezonskost, cikli in neregularna komponenta, pri čemer uporabimo aditivni ali pa multiplikativni model [28]:

- Aditivni model: trend + sezonskost + cikli + neregularna komponenta.
- Multiplikativni model: trend * sezonskost * cikli * neregularna komponenta.

Na sliki 2 je prikazana časovna vrsta, ki je bila razdeljena na tri komponente, ki so trend, sezonskost in ostanek ali neregularna komponenta, pri čemer je bil uporabljen aditiven model. V nadaljevanju bomo predstavili pomen teh komponent.



Slika 2: Časovna vrsta, ki je bila razčlenjena na trend, sezonskost in ostanek.

3.3.1 Trend

Trend časovne vrste predstavlja dolgoročno gibanje povprečja časovne vrste [28]. Trend je najdaljša komponenta celotne časovne vrste in ima lahko različne vrste gibanja, na podlagi oblike, ki le-ta zavzame [14]. Trend je namreč linearen, ali pa tudi kvadraten, kubičen, ali na katerikoli n-ti polinom. Kot lahko opazimo v sliki 2, je trend linearen, ker sledi linearnem gibanju.

3.3.2 Sezonskost

Sezonskost je izraz, ki se nanaša na ponavljajoča se gibanja, ki se ponavljajo v rednih časovnih presledkih [13]. Primer sezonskosti je 10-letno gibanje prodaje sladoledov, pri čemer upoštevamo dnevno gibanje. V tem primeru bomo opazili povečano prodajo v poletnem času in nato upad v zimskem obdobju, ki bo tvorila sezonskost tovrstne časovne vrste. Na sliki 2 opazimo letno sezonskost, saj se gibanje vsako leto poveča in zmanjša po določnem vzoru.

3.3.3 Cikli

Cikli se nanašajo na vrednosti časovne vrste, ki so odvisne od določenega števila prejšnjih vrednosti le-te [12]. Cikli namreč nimajo stalnega gibanja, kot je to v primeru trenda, ali sezonskosti, temveč je naslednja vrednost povezana le z določenim številom prejšnjih vrednosti [12].

3.3.4 Ostanek

Ostanek je razlika med opazovano vrednostjo in z modelom ocenjeno vrednostjo [28]. Ostanek je prikazan na sliki 2, pri čemer opazimo vrednosti, ki jih model ne uspe zajeti in zato postanejo ostanek.

3.4 Strojno učenje

Strojno učenje je veja umetne inteligence, ki računalniškim sistemom omogoča neposredno učenje na podlagi primerov, podatkov in izkušenj, kar omogoča učenje iz podatkov in ne po vnaprej programiranih pravilih [8]. Strojno učenje se uporablja na številnih področjih, ki gredo od odkrivanja razreda določenih instanc do napovedovanja določenega dogodka, kot so vremenske razmere, ali celo napovedovanje vrednosti delnice kot v primeru te zaključne naloge. Strojno učenje temelji na določenem številu označenih ali neoznačenih primerov, pri čemer stolpci označujejo atribute. Poleg tega je strojno učenje razdeljeno v dve glavni skupini, ki sta nadzorovano in nenadzorovano učenje. Glavna razlika med njima je prisotnost razreda, pri čemer je pri nadzorovanem učenju prisoten, pri nenadzorovanem pa ne. Ti dve glavni skupini se nadaljnje podrobneje delita na regresijo, klasifikacijo, asociacijo in razvrščanje v skupine. Prvi dve spadata v skupino nadzorovanega učenja in bosta opisani v naslednjih razdelkih, medtem ko drugi dve spadata v skupino nenadzorovanega učenja.

3.4.1 Klasifikacija

Klasifikacija je vrsta nadzorovanega učenja, pri čemer napovedujemo razred primerov. V naslednjem razdelku bomo opisali enega od najbolj uporabljenih algoritmov strojnega učenja, ki temelji na logistični funkciji, to je logistična regresija.

Logistična regresija

Logistična regresija je algoritem strojnega učenja, ki klasificira primere na podlagi logistične funkcije oziroma sigmoidove funkcije, ki je prikazana v naslednji enačbi [34]:

$$Sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

Sigmoidova funkcija je na y-osi omejena med 0 in 1 in ima obliko črke S [34]. V primeru logistične regresije bomo kot argument sigmoidne funkcije uporabili linearno kombinacijo atributov strojnega učenja, kot je prikazana v formuli 3.2, pri čemer so $A_1, A_2, ..., A_n$ atributi strojnega učenja in x_0, ..., x_n uteži modela [34].

$$P(x) = \frac{1}{1 + e^{-(x_0 + x_1 A_1 + x_2 A_2 + ... + x_n A_n)}} \tag{3.2}$$

Ker je sigmoidova funkcija omejena med 0 in 1, bo tako za formulo tudi prikazano v 3.2, ker nam omogoča, da uporabimo to za prikaz verjetnosti pripadanja primera določnemu razredu [34].

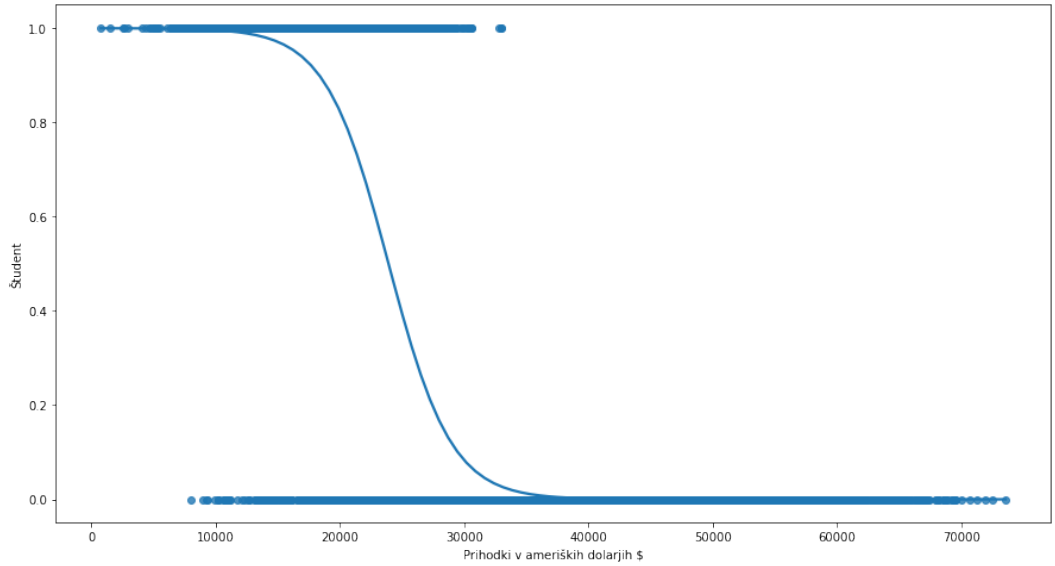
Na sliki 3 je prikazan primer logistične regresije, pri čemer je upoštevana dvovrednostna razredna spremenljivka in je cilj razvrščati osebe na študente in ne-študente na podlagi prihodka. Na y-osi imamo verjetnost, da ena oseba spada v določeno skupino, na x-osi pa njen prihodek. Kot lahko opazimo na sliki 3, uporabimo logistično funkcijo, ki je bila izračunana na podlagi primerov, ki so bili uporabljeni v testni množici algoritma strojnega učnega logistična regresija.

3.4.2 Regresija

Regresija je tako kot klasifikacija vrsta nadzorovanega učenja, pri čemer se napoveduje številčna vrednost primera. V naslednjih razdelkih bomo opisali najpomembnejše algoritme strojnega učenja, ki temeljijo na regresiji.

Linearna regresija

Linearna regresija je eden od preprostejših in uporabnih algoritmov strojnega učenja in je primerna, kadar podatki kažejo na linearno odvisnost. Linearna regresija temelji na iskanju linearnih kombinacijah atributov strojnega učenja s pomočjo razreda. Namreč,



Slika 3: Primer logistične regresije, pri čemer klasificiramo študente na podlagi prihodka. Na sliki je prikazana tudi sigmoidova funkcija, ki nam omogoča klasifikacijo primerov.

če so A_1, ..., A_n atributi strojnega učenja in so A_1^i, ..., A_n^i posamezne vrednosti atributov pri i-tem primeru, bo linearna regresija izračunala uteži x_0, ..., x_n, ki oblikujejo naslednjo enačbo, pri čemer je y_i napovedana vrednost [33]:

$$y_i = x_0 + x_1 A_1^i + x_2 A_2^i + ... + x_n A_n^i \tag{3.3}$$

Ena od značilnosti linearne regresije je, da se lahko poleg linearne funkcije prilega vsaki polinomski funkciji [14].

SVR

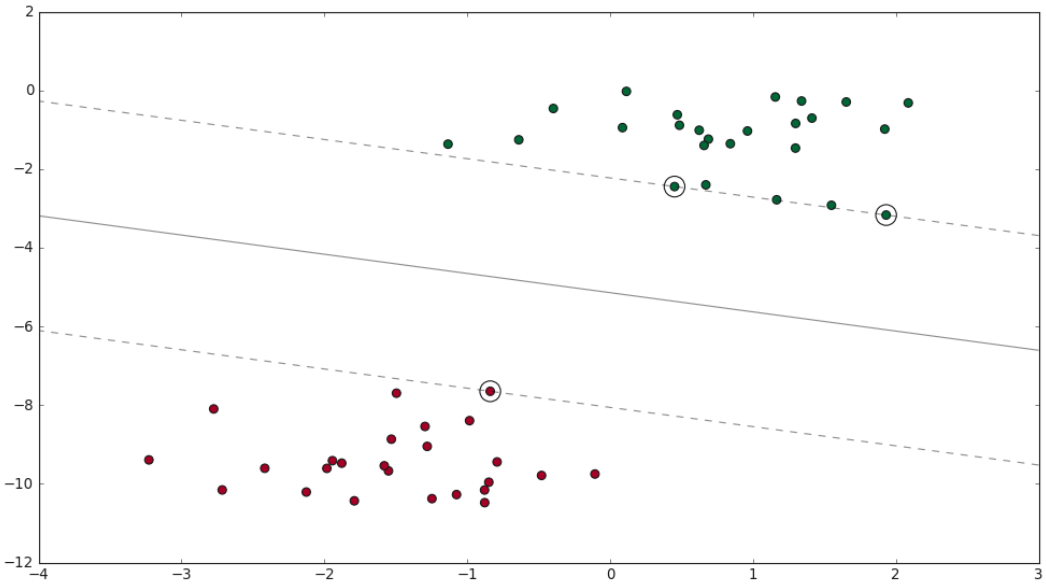
SVR (ang. Support vector regression – regresijska metoda podpornih vektorjev) je algoritem strojnega učenja, ki deluje na podobnem principu kot SVM (ang. Support vector machine – metoda podpornih vektorjev). Namreč, pri algoritem strojnega učenja SVM v primeru testne množice v obliki:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \in \mathbb{R}^n \times -1, 1 \tag{3.4}$$

pri čemer -1, 1 prikazujejo pripadnost primera določenem razredu, bo algoritem postavil hiperravnine (ang. hyperplanes), ki razdelijo primere na podlagi pripadnosti določenem razredu [10].
Po postavitvi prve hiperravnine bodo postavljeni tudi razponi na podlagi vektorjev,

imenovani podporni vektorji [10]. Za izračun hiperravnine pa uporabimo različne linearne ali nelinearne pristope na podlagi podatkov, ki jih imamo na voljo. Napoved pa bo temeljila le na določnem število primerov testne množice [10].
Princip delovanja SVR je zelo podoben. Namreč, pri regresiji podatkov namesto iskanja hiperravnine, ki razdeli testno množico, SVR uvaja funkcijo izgube ϵ-občutljiva za izračun hiperravnine, tako da imajo napovedane vrednosti učnih vzorcev največ ϵ odstopanje od dejanskih vrednosti [38]. Hiperravnina in ϵ določata ϵ-neobčutljivo cev (ali pas) za izračun posplošitvenih mej za regresijo [38].

Na sliki 4 je prikazan primer SVM, pri čemer imamo primere, ki so pobarvani glede na pripadnost razreda. Kot lahko opazimo v sredini je z neprekinjeno črto prikazana hiperavnina, ki razdeli primere vsakega posameznega razreda. Črtkane črte pa predstavljajo razpone. Primeri, ki se nahajajo na le-teh in so obkroženi pa so podporni vektorji [29].



Slika 4: Primer SVM, pri čemer so prikazani posamezni primeri v obliki točk, hiperravnina, razponi in podporni vektorji [29].

3.5 Metrike uspešnosti

Metrike uspešnosti merijo napako modelov strojnega učenja. V tem razdelku bodo predstavljene metrike uspešnosti za klasifikacijske in regresijske modele strojnega učenja.

3.5.1 Klasifikacijske metrike uspešnosti

Za meritev napak pri klasifikaciji z dvovrednostnimi razredni spremenljivki, primerjamo napovedane vrednosti modela z dejanskimi vrednosti razreda, ki namreč tvorijo naslednje kategorije rezultatov [31]:

- TP: Delež vrednosti, ki so pozitivne in so bile pravilno klasificirane.
- TN: Delež vrednosti, ki so negativne in so bile pravilno klasificirane.
- FP: Delež vrednosti, ki so negativne in niso bile pravilno klasificirane.
- FN: Delež vrednosti, ki so pozitivne in niso bile pravilno klasificirane.

Na podlagi teh kategorij so v spodnjih razdelkih predstavljene metrike uspešnosti.

Klasifikacijska točnost

Klasifikacijska točnost (ang. Accuracy) je metrika uspešnosti, ki predstavlja delež vrednosti, ki so bile klasificiranje na pravilen način in se izračuna na podlagi formule 3.5:

$$Točnost = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

Preciznost klasifikacijskega modela

Preciznost klasifikacijskega modela (ang. Precision) je metrika uspešnosti, ki predstavlja delež pozitivnih pravilno klasificiranih vrednosti med vsemi pozitivno kvalificiranimi vrednosti:

$$Preciznost = \frac{TP}{TP + FP} \tag{3.6}$$

Priklic klasifikacijskega modela

Priklic klasifikacijskega modela (ang. Recall) je metrika uspešnosti, ki predstavlja delež pozitivnih pravilno klasificiranjih vrednosti med vsemi vrednosti, ki so pozitive. Priklic predstavlja zelo pomembno metriko, ko je cilj imeti čim manjši delež lažnih pozitivnih vrednosti, torej vrednosti, ki so pozitivne, vendar so bile klasificirane kot negativne. Primeri, kjer je priklic pomemben, so na primer farmacevtski testi, kot na primer testi za nosečnost.

$$Priklic = \frac{TP}{TP + FN} \tag{3.7}$$

Mera F1

Mera F1 predstavlja harmonično povprečje med natančnostjo in priklicem [32]:

$$F1 = 2 \cdot \frac{Natančnost \cdot Priklic}{Natančnost + Priklic} \tag{3.8}$$

3.5.2 Regresijske metrike uspešnosti

V nadaljevanju bomo pregledali metrike uspešnosti, ki se uporabljajo za ocenitev napak, ki jih regresijski model naredi pri napovedovanju številčne vrednosti.

RMSE

RMSE (ang. Root Mean Squared Error – koren povprečne kvadratne napake) je eden od glavnih absolutnih metriki uspešnosti pri regresiji, ki kaznuje velike napake pri napovedovanju. Enačba 3.9 prikazuje formulo za izračun RMSE-ja, pri čemer je:

- n je skupno število podatkovnih točk.
- x_1, ..., x_n dejanska vrednost točke.
- $\overline{x}_1$, ..., $\overline{x}_n$ napovedana vrednost točke.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\bar{x}_i - x_i\right)^2} \tag{3.9}$$

Koeficient determinacije – R^2

Koeficient determinacije ali R^2, v primerjavi z RMSE, je relativna matrika uspešnosti, ki prikazuje napako regresijskega modela [30]. Namreč, koeficient determinacije je statistično merilo, kako dobro se regresijske napovedi približujejo dejanskim podatkovnim točkam in, ko je enak 1, pomeni, da se regresijske napovedi popolnoma ujemajo s podatki [30]. V 3.10 je prikazana formula za izračun R^2, pri čemer je:

- n je skupno število podatkovnih točk.
- y_i dejanska vrednosti točke.
- f_i napovedana vrednosti točke.
- $\overline{y}$ aritmetično povprečje vseh točk [30].

$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}(y_i - f_i)^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.10}$$

3.6 Python

Python je programski jezik, ki predstavlja enega od glavnih jezikov za podatkovno znanost in je zaradi svoje priročnosti postal najbolj priljubljen jezik za strojno učenje in umetno inteligenco [3]. Naša programska rešitev je bila v celoti zasnovana v programskem jeziku Python.

3.7 Zajemanje internetnih virov

Zajemanje internetnih virov (ang. web harvesting oziroma web scraping) je aktivnost, ki se osredotoča na "zajemanje" oziroma prenos podatkov od določenega spletnega mesta, kot so družbena omrežja za nadaljnjo obdelavo, ali analizo podatkov.

4 Predštudija

Preliminarna študija na področju napovedovanja cen delnic je najprej zajemala dogodke, ki vplivajo na vrednost delnic, kamor se uvrščajo različni dejavniki, vključno s pričakovanimi ali nepričakovanimi dogodki. Torej, v tem poglavju se bomo osredotočali na prvo raziskovalno vprašanje: Ali je gibanje razpoloženja na družbenih omrežjih po nekemu dogodku povezano z gibanjem vrednosti borznega indeksa države, v kateri se je dogodek zgodil?

V raziskavi [27] avtorjev Tavor T. in drugi dokazali vpliv nesreč in terorizma na delniškem trgu. Zbirali so podatke treh vrst dogodkov: umetne nesreče, naravne nesreče in teroristične napadi, ki so se zgodili v različnih državah po svetu. Poleg tega so uporabili podatke o glavnih finančnih indeksih vsake države, v kateri so se ti dogodki zgodili. Rezultati razkrivajo, da se med naravnimi nesrečami borzni indeks države, v kateri se je dogodek zgodi, zmanjša na dan dogodka in v naslednjih dveh dneh. Zato bi morali vlagatelji na dan nesreče prodati indeks in ga zadržati dva dni. Nasprotno pa se med umetnimi nesrečami, ali terorističnimi napadi indeks zniža samo na dan dogodka in naslednji dan, zato bi morali vlagatelji na dan nesreče prodati indeks in ga držati do konca prvega delovnega dne po dogodku.
Na podlagi tovrstne raziskave smo razmislili o mogoči razširitvi te tematike na področju družbenega omrežja Twitter. Namreč, določen dogodek povzroči tudi, da uporabniki začnejo objavljati misli in informacije o tem dogodku in na podlagi teh objav pa lahko sklepamo padec ali rast vrednosti indeksa države, v kateri se je ta dogodek zgodil.

4.1 Zajem podatkov

Da bi to razširili na širše področje, in sicer z uporabo analize sentimenta v družbenih omrežjih in vpliva dogodkov za napovedovanje vrednosti delnic, smo začeli zbirati podatke o dogodkih in nesrečah v državah, v katerih je državni in uradni jezik angleščina, ker nam bi omogočila lažjo obdelavo in razumevanje podatkov in enostavnejši zajem podatkov iz družbenih omrežji. Zato smo izbrali državo Združeno kraljestvo, ker izpolnjuje zgoraj navedene pogoje, poleg tega pa ima borzni indeks države FTSE, ki zajema določeno število podjetji znotraj Združenega kraljestva. Na podlagi tega smo

začeli zbirati podatke o nesrečah in dogodkih v Združenem kraljestvu, pri čemer smo izbrali podatkovno zbirko Epcresilience, ki vsebuje dogodke iz celega sveta, vendar je predvsem osredotočena na dogodke v Združenem kraljestvu [9].

4.2 Predobdelava in analiza podatkov

Podatkovna zbirka Epcresilience vsebuje dogodke iz vsega sveta, zato smo morali filtrirati dogodke, ki so se zgodili v Združenem kraljestvu. Na koncu tega koraka je zbirka je vsebovala skupno 227 dogodkov v obdobju med letoma 1864 in 2019. Ker je bil Twitter ustanovljen leta 2006, smo se odločili, da se osredotočimo le na obdobje od leta 2010 do vključno leta 2019. Za tovrstna leta podatkovna zbirka vsebuje skupaj 36 dogodkov, pri čemer so vključeni teroristični napadi, eksplozije, letalske nesreče, izbruh bolezni, streljanja z orožjem in ekstremna vremenska razmerja, ki so trajale nekaj dni, kot so poplave in drugo. Kot lahko opazimo, tudi če vzamemo 10-letno obdobje, je dogodkov precej malo in le določen delež dogodkov je silovitih, da bi imeli konkreten učinek na državno ekonomijo.

4.3 Zaključek

Pri prejšnjem razdelku 4.2 smo opazili, da ima Združeno kraljestvo precej malo nesreč in negativnih dogodkov, ki bi lahko vplivali na vrednosti indeksa države. V našem primeru bi moral časovni obseg objav zajemati daljše časovno obdobje, da bi zajemali objave na omrežju Twitter in naredili analizo razpoloženja teh objav. Poleg tega pa se z ročno primerjavo vrednost indeksa države po nesreči, ali dogodku v večini primerov vidi, da dogodki nimajo vpliva na borzni indeks države. Zaradi omenjenih težav je bila ta zasnova opuščena za trenutno zaključno nalogo.

4.3.1 Povezava med večjimi dogodki in gibanje delnice

Kot je opisano v zgornjih razdelkih poglavja o predštudiji vidimo, da rezultati niso pokazali obstoja povezave med večjimi dogodki in gibanjem delnic. Zaradi tega smo ovrgli prvo raziskovalno vprašanje: Ali je gibanje razpoloženja na družbenih omrežjih po nekemu dogodku povezano z gibanjem vrednosti borznega indeksa države, v kateri se je dogodek zgodil?

5 Metodologija glavne študije

V tem poglavju obravnavamo metodologijo, uporabljeno v zaključni nalogi. V naslednjih razdelkih bomo predstavili korak pridobivanja podatkov o delnici in tvitih, način interpolacije manjkajočih vrednosti delnice, obdelavo podatkov in načine ustvarjanja dodatni podatkov na osnovi obstoječih podatkov, ki bodo pozneje uporabljeni kot atributi strojnega učenja. V poglavju bo predstavljen tudi korak analize razpoloženja, pri čemer bodo predstavljena orodja in rezultati le-tega.

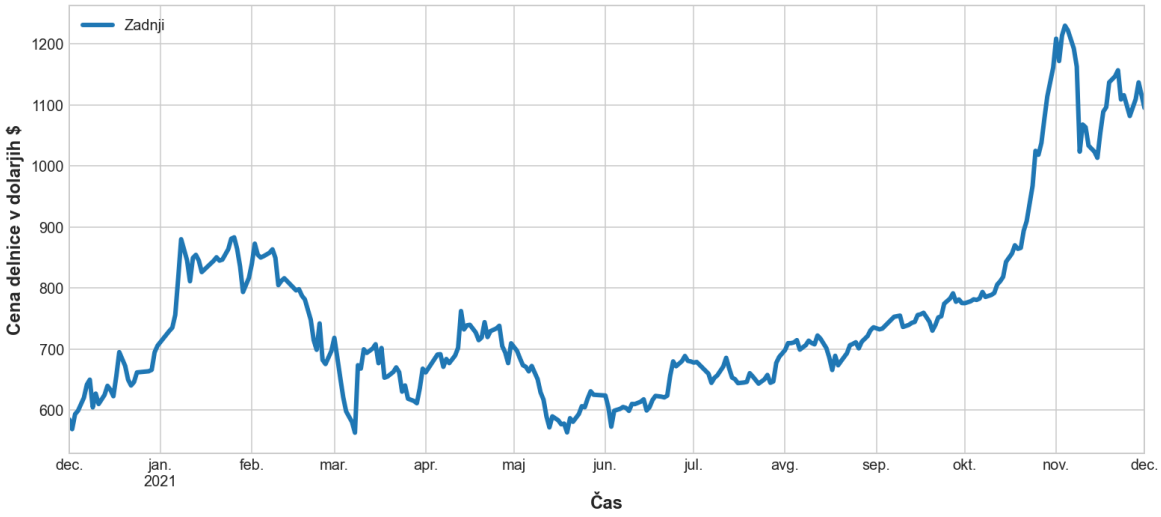
5.1 Podatki o delnici

Za izbiro delnice podjetja smo se odločil na podlagi raziskave avtorjev Bing L. idr. [6]. Odločili smo se, da izberemo podjetje, ki se uvršča med IT, medijska in finančna podjetja, ker so bolj občutljiva na razpoloženje na družbenih omrežjih. Poleg tega smo morali izbrati podjetje, o katerem se veliko razpravlja na družbenih omrežjih, cena delnice pa je v zadnjem letu izrazito nihala. Na koncu smo se odločili, da bomo obravnavali delnico TSLA podjetja Tesla, ki izpolnjuje vse zgoraj navedene pogoje.
Izbrali smo obdobje med 1. decembrom 2020 in 1. decembrom 2021. Kot je razvidno iz slike 5, to obdobje zajema veliko gibanje vrednosti delnice, saj se je cena gibala od približno 600 dolarjev konec leta 2020 do več kot 1000 dolarjev proti koncu leta 2021.

5.1.1 Pridobivanje podatkov

Knjižnica yfinance

Vrednosti zadnjega tečaja in preostale podrobnosti delnice TSLA so bile pridobljene z uporabo knjižnice yFinance za programski jezik Python. Yfinance je odprtokodna knjižnica, ki omogoča prenos podatkov o določeni delnici iz strežnikov Yahoo! Finance [25]. Z uporabo zgoraj navedene knjižnice smo prenesli podatke delnice TSLA. Tako dobljena tabela vsebuje stolpce datum, odpiralni, zadnji, najvišji, najnižji tečaj in obseg trgovanja delnice za vsak posamezen trgovalni dan za obdobje od 1. decembra 2020 do 1. decembra 2021, brez praznikov in drugih dela prostih dni, ko je borza zaprta. Podatki so bili shranjeni v obliki CSV (ang. Comma-separated values –



Slika 5: Vrednosti zadnjega tečaja delnice TSLA v valuti ameriški dolar v obdobju od 1. decembra 2020 do 1. decembra 2021.

vrednosti, ločene z vejico).

5.1.2 Predobdelava podatkov

Interpolacija vrednosti delnice

TSLA kotira na newyorški borzi NASDAQ, ki je odprta le določene dni v tednu. Namreč, borza je ob sobotah, nedeljah in praznikih zaprta, zato vrednosti delnic ob teh dneh ne obstajajo. Po drugi strani so tviti objavljeni vsak dan, ne glede na uro dneva, ali dan v tednu. Ker algoritmi strojnega učenja, ki bodo uporabljeni pozneje, pri napovedi vrednosti delnice, ne podpirajo manjkajočih vrednosti, smo morali izračunati le-te. Da smo lahko to naredili, smo uporabili tehniko, ki je bila uporabljena v raziskavi [21] avtorjev Sasank Pagolu V. in drugi. Če poznamo vrednosti x in y in imamo med njima neznane vrednosti, lahko dan po x aproksimiramo z enačbo (x + y) / 2 in to ponavljamo rekurzivno, dokler ne zapolnimo vseh vrednosti [21].

Dodatni atributi

Poleg samih podatkov delnice smo dodali dodatni stolpec, ki so ga prav tako uporabili tudi avtorji Kordonis J. in drugi v raziskavi [18], ki se zaračuna na podlagi naslednje enačbe, pri čemer se izraza ”zadnji” in ”odpiralni” nanašata na zadnji tečaj oziroma odpiralni tečaj.

$$Odstotek\ spremenitve = \frac{zadnji - odpiralni}{odpiralni}$$

Ta enačba prikazuje, koliko se je vrednost delnice spremenila med odpiralnim tečajem in zadnjim tečajem delnice.

5.2 Podatki o tvitih

Po določitvi delnice smo morali določiti tudi družbeno omrežje, na podlagi katerega bomo naredili analizo razpoloženja objav. Nazadnje smo se odločili za socialno omrežje Twitter, saj vsakemu uporabniku omogoča , da izrazi svoje stališče o določenih tematikah s sporočili, dolgimi največ 280 znakov, pri čemer lahko objavi dodatne večpredstavnostne vsebine, kot so slike, ali posnetki. Na ta način lahko analiziramo razpoloženje uporabnika na podlagi sporočila, ki ga je objavil.

5.2.1 Pridobivanje podatkov

Knjižnica Snscrape

Za pridobivanje tvitov smo uporabili orodje Snscrape. Snscrape je odprtokodna knjižnica za zajemanje internetnih virov (ang. Scraper) za različna družbena omrežja [17]. S tem orodjem smo prenesli tvite, povezane s podjetjem TESLA in delnicami le-te imenovana $TSLA. Na podlagi raziskave avtorjev Skuza M. [4] smo se odločili, da bomo iskali le tvite, ki se izrecno sklicujejo na delnico TSLA, in sicer tvite, ki vsebujejo besedo "TSLA", ali "$TSLA". Poleg samega filtriranja besed smo filtrirali in odstranili tudi tvite, ki so vsebovali povezave do spletnih strani, torej tvite, ki so vsebovali niza "http" ali "www". S tem smo se izognili nezaželenim objavam (ang. spam), ki bi lahko vplivale na korak sentimentalne analize. Hkrati smo odstranili tudi tvite, ki niso bili v angleškem jeziku. Preneseni očiščeni tviti zajemajo obdobje med 1. decembrom 2019 in 1. decembrom 2021 in obsegajo skupaj 1.234.137 tvitov (325MB), vključujejo polja datum in čas, identifikacijska številka tvita, vsebina tvita, povezava do tvita, število všečkov tvita, število retvitov, število odgovorov, število sledilcev uporabnika in število prijateljev uporabnika. Podatki so bili shranjeni v obliki CSV.

5.2.2 Predobdelava podatkov

Nastavitev enotnega časovnega pasu

Preneseni tviti so shranjeni v univerzalnem koordinatnem časovnem pasu 00:00 (UTC 00:00). Po drugi strani TSLA kotira na newyorški borzi NASDAQ, ki je v vzhodnem standardnem časovnem pasu (UTC−05:00) [19]. Ker je naš cilj uporabiti razpoloženje

uporabnikov pri napovedi vrednosti delnice, smo vse tvite nastavili na vzhodni standardni časovni pas, ker nam omogoča pravilno klasifikacijo tvitov za vsak posamezni dan.

VADER

Eden najpomembnejših delov je bila izbira najboljšega algoritma za prepoznavanje sentimentov v tvitih. Na voljo je kar nekaj alternativ, kot so NLTK (ang. Natural Language Toolkit – orodje za naravni jezik), VADER (ang. Valence Aware Dictionary and sEntiment Reasoner – valenčno ozaveščen slovar in utemeljitve čustev), SentiWordNet, TextBlob in drugi. V raziskavi [5] avtorjev Sohangir S. in drugi so primerjali pristope sentimentalne analize, ki temeljijo na strojnem učenju, s pristopi, ki temeljijo na leksikonu (ang. Lexicon based approches) na finančnih tvitih iz družbenega omrežja StockTwit. Pristopi strojnega učenja so temeljili na logistični regresiji SVM in Naive Bayes, medtem ko so bili pristopi, ki so temeljili na leksikonu VADER, SentiWordNet in TextBlob. Med vsemi se je VADER izkazal za najboljšega, saj je s 94-odstotno natančnostjo razvrstil tvite na negativne in pozitivne. Poleg tega pa je bil hitrejši od pristopov, ki so temeljili na strojem učenju. Zaradi teh prednosti smo VADER izbrali kot knjižnico za izvajanje analize čustev na tvitih.

Knjižnica za analizo razpoloženja VADER temelji na leksikonu, kar pomeni, da temelji na nenadzorovanem učenju in vhodni podatki ne vsebujejo razreda. Poleg tega, da temelji na leksikonu pa je tudi pristop, ki deluje na podlagi pravil (ang. rule-based) [16]. Pristopi, ki temeljijo na pravilih, oblikujejo slovnična in logična pravila za dodeljevanje čustev in uporabijo leksikon za dodeljevanje čustev, ali polarnosti besedam [20].

VADER po analizi določenega besedila vrne štiri komponente, ki predstavljajo rezultate analize razpoloženja [16]:

- Negativno – predstavlja delež besedila, ki ima negativno razpoloženje.
- Nevtralno – predstavlja delež besedila, ki ima nevtralno razpoloženje.
- Pozitivno – predstavlja delež besedila, ki ima pozitivno razpoloženje.
- Kombinirana ocena (ang.compound) – se izračuna s seštevanjem valenčnih ocen vsake besede v leksikonu, ki se prilagodi v skladu s pravili in nato normalizira tako, da je med –1 (najbolj negativno) in +1 (najbolj pozitivno) [16].

Kombinirana ocena je tudi metrika za klasificiranje besedila. Če je ocena manjša od –0,05, lahko besedilo razvrstimo kot negativno, če je kombinirana ocena večja od 0,05, je besedilo razvrščeno kot pozitivno, v vseh drugih primerih pa je besedilo nevtralno [16].

Priprava podatkov za analizo razpoloženja

Za pripravo tvitov za fazo analize čustev smo odstranili nepotrebne komponente stavkov in očistili tvite, da bi pripomogli k čim boljšemu prepoznavanju sentimenta v tvitih.

Prvi korak je zajemal uporabniška imena. Uporabniška imena v tvitih so na začetku označena s simbolom @, niz, ki sledi pa je uporabniško ime. Uporabniško ime ne vsebuje nobenih podrobnosti za ugotavljanje razpoloženja tvitov, poleg tega pa lahko razveljavi analizo razpoloženja tvitov – na primer uporabnik z uporabniškim imenom IamHappy, zato so bili uporabniki s takimi imeni odstranjeni.

Sledeči korak se je nanašal na simbol #, ki je na Twitterju zapisan na začetku določenih besed, ki označujejo temo tvita. Na primer uporabnik, ki tvita o delnici TSLA, lahko zapiše #TSLA. Zaradi lažjega prepoznavanja razpoloženja v tvitih, je bil simbol # odstranjen iz besed, ki se začnejo s tem simbolom – na primer beseda #abc je bila pretvorjena v abc.

Nato je celotna vsebina tvita zaradi lažje obdelave besedila postavljena v eno vrstico, vsi simboli HTML (ang. Hypertext markup language – jezik za označevanje nadbesedila) pa so bili prevedeni v dejanske znake (na primer HTML znak &amp je bil preveden v simbol &). V naslednjem koraku so bile odstranjene tudi številke, saj ne dodajajo nobenih podrobnosti za zaznavanje razpoloženja tvitov. Nazadnje smo besedilo tokenizirali, torej smo tvit razdelili na posamezne nize in z lematizacijo besedila pretvorili besede v njihovo osnovno obliko. Na koncu smo tvite ponovno sestavili v posamezna besedila.

Preoblikovanje besedila ni vključevalo presledkov, znakov in velikih tiskanih črk, saj VADER ne analizira le pomen besed, temveč preverja tudi pomen nekaterih simbolov v besedilu. Simbola, kot sta :), ali :D in zlasti prisotnost ločil, kot so !, ? ter prisotnost velikih tiskanih črk vplivajo na zaznano moč čustev tvita. [16].

5.2.3 Analiza razpoloženja tvitov

Kot je bilo povedano v razdelku 5.2.2 je bil kot orodje za sentimentalno analizo izbran VADER. S tem orodjem smo besedilo vsakega tvita analizirali in pridobili vrednosti razpoloženja vsakega posameznega tvita, ki je negativna, nevtralna, pozitivna in kombinirana ocena. Na podlagi naših namenov smo med vsemi obdržali le kombinirano oceno, ker najbolj prikazuje razpoloženje besedila. Na podlagi te ocene smo tudi kla-

sificirali tvite na negativne, nevtralne in pozitivne na osnovi pravil predstavljenih v razdelku 5.2.2.

5.3 Uporaba podatkov

V naslednjih razdelkih bomo tukaj predstavljene podatke uporabili za nadaljnjo analizo. V naslednjem poglavju bomo iskali način združevanja razpoloženja tvitov v enotne parametre, ki jih lahko uporabimo pri razumevanju povezave med delniškim trgom in razpoloženjem objav. Na osnovi tega pa bomo sestavili napovedovalne modele, ki bodo predstavljeni v poglavjih 7 in 8

6 Povezava med razpoloženjem v tvitih in delnico TSLA

V tem poglavju bomo predstavili povezavo med razpoloženjem v tvith in delnico TSLA, kar je tudi eno od raziskovalnih vprašanj. V tem poglavju se bomo osredotočali na raziskovalno vprašanje: Ali je gibanje razpoloženja na omrežju Twitter povezano z gibanjem vrednosti delnice TSLA? V prvem razdelku bomo predstavili koncept Pearsonove korelacije in na podlagi le-te primerjali različne strategije agregiranja, ki so nam omogočile združevanje našega nabora tvitov v enotne parametre, ki imajo v našem primeru frekvenco enkrat dnevno, kot tudi vrednost zadnjega tečaja delnice. Na koncu bodo predstavljene korelacije med vsemi atributi, ki izhajajo iz vrednosti delnice in tvitov.

6.1 Pearsonov korelacijski koeficient

Pearsonov korelacijski koeficient je matematična in statistična številska mera, ki predstavlja velikost linearne povezanosti med dvema spremenljivkama X in Y, merjenih na istem predmetu preučevanja [35]. Koeficient je definiran kot razmerje med kovarianco in produktom obeh standardnih odklonov [35]. Formula za izračun Pearsonovega koeficienta korelacije P je prikazan v formuli 6.1, pri čemer so:

- $Cov(X, Y)$ – kovarianca spremenljivke X in Y,
- σ_X – standardni odklon spremenljivke X,
- σ_Y – standardni odklon spremenljivke Y.

$$P = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{6.1}$$

Vrednosti, ki Pearsonov korelacijski koeficient lahko zavzamejo, se gibljejo v intervalu med [-1,1]. Negativne vrednosti namreč predstavljajo negativno korelacijo med dvema spremenljivkama, pri čemer je največja negativna korelacija enaka –1. Po drugi strani pa pozitivne vrednosti predstavljajo pozitivno korelacijo med dvema spremenljivkama in je največja pozitivna korelacija enaka 1. Vrednost 0 pa označuje najnižjo korelacijo in prikazuje, da korelacija med dvema spremenljivkama ne obstaja.

6.2 Agregacijske strategije

Cene delnic se gibljejo le v določenih urah v določenih dneh, zato imajo za določen dan le eno odpiralno in eno zapiralno vrednost. Po drugi strani so tviti objavljeni ne glede na uro, ali dan, količina tvitov v določenem času pa se spreminja. Ker je naš cilj napoved zadnjega tečaja delnice za vsak posamezni dan, smo morali opredeliti strategijo za združevanje tvitov v danem intervalu v enotne vrednosti, ki jih bomo pozneje uporabili za določitev korelacije med delniškim trgom in tvitih ter nazadnje za izgradnjo napovedovalnega modela strojnega učenja.

Pri predobdelavi tvitov smo dodali dva stolpca namenjena razpoloženju vsakega posameznega tvita, ki so klasifikacija tvita (negativen, nevtralen in pozitiven) in kombinirana ocena (med [-1, 1]), ki so bili predstavljeni v poglavju 5.2.3. Na podlagi teh dveh vrednosti so bile definirane naslednje vrednosti, ki bodo izračunane na podlagi upoštevanega intervala:

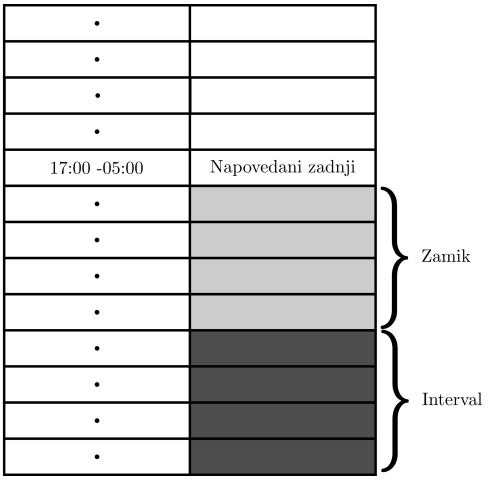
- Odstotek negativnih tvitov v danem intervalu.
- Odstotek nevtralnih tvitov v danem intervalu.
- Odstotek pozitivnih tvitov v danem intervalu.
- Agregacija kombiniranih ocen tvitov v danem intervalu.

Prva tri atributa sta izpeljana iz klasifikacije tvitov, medtem ko je četrti atribut izpeljan iz kombinirane ocene tvitov.

6.2.1 Določitev optimalnega zamika in intervala

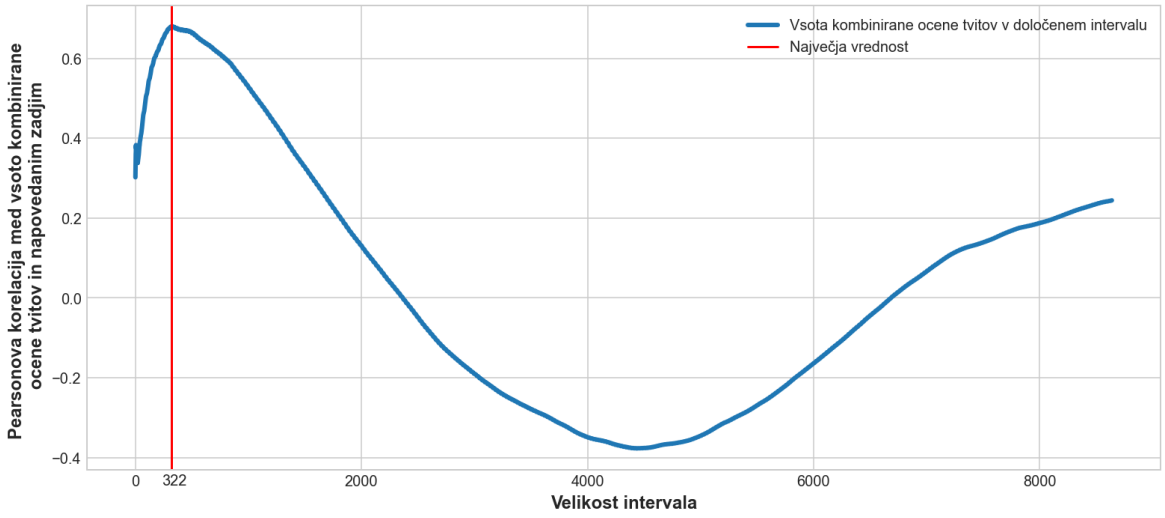
Po določitvi atributov je bilo treba določiti tudi velikost intervala in velikost odmika od napovedovanega zadnjega tečaja delnice, kot je razvidno na sliki 6. Po določitvi intervala in zamika bomo lahko opredelili atribute, ki nam omogočajo prehod s spremenljive dolžine na fiksno dolžino, in sicer enkrat dnevno.

Ker je skupna ocena razpoloženja ključnega pomena, smo le-to uporabili tudi pri postopku določanja intervala in iskali največjo korelacijo med agregirano vrednostjo velikosti intervala in vrednosti napovedanega zadnjega tečaja delnice. Določili smo fiksni zamik 10 ur od zadnjega tečaja, pri čemer smo spreminjali velikost intervala od 1 ure do 8640 ur (360 dni), da bi razumeli, pri kateri vrednosti intervala obstaja največja korelacija z napovedanim zadnjim tečajem. Pri tem smo primerjali različne strategije agregiranja tvitov, ki so vsota, povprečje in produkt. Tako dobljene vrednosti smo



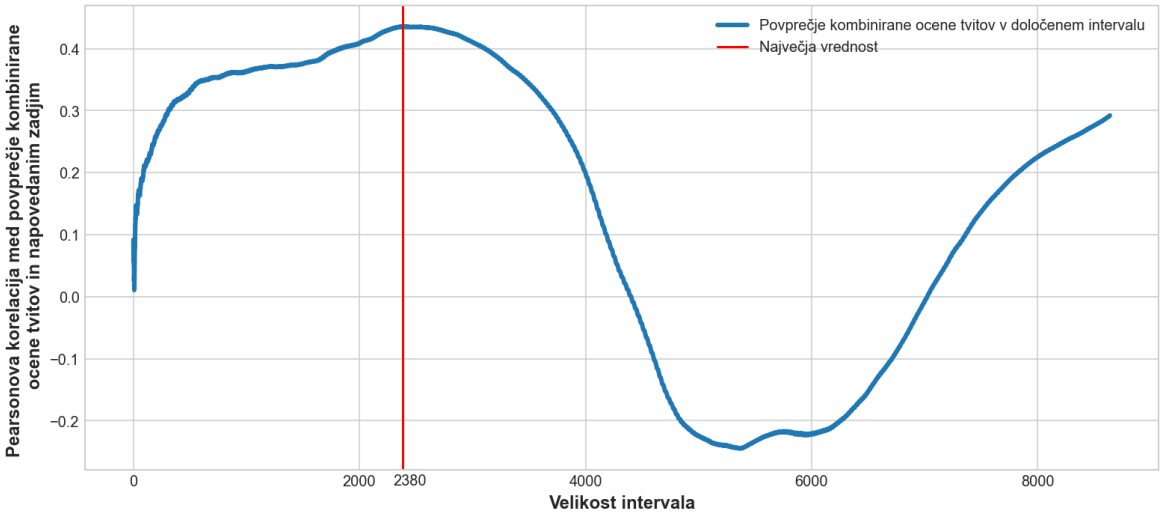
Slika 6: Prikaz komponente zamika in intervala od zadnjega tečaja, ki je bila uporabljena pri izračunu Pearsonovega korelacijskega koeficienta.

primerjali z napovedanim zadnjim tečajem določenega dne z uporabo Pearsonove korelacije.
Kot je razvidno iz slike 7, se je pri uporabi Pearsonove korelacije kot najboljša strategija agregacije izkazala vsota v primerjavi s povprečjem na sliki 8 in produktom. Kot je razvidno iz slike 7, ima interval velikosti 322 ur največjo korelacijo z napovedanim zadnjim tečajem delnice, pri čemer je le-ta 0,66.



Slika 7: Odvisnost Pearsonove korelacije med napovedanim zadnjim tečajem in velikostjo intervala vsote kombinirane ocene.

Nato je bilo treba določiti tudi optimalnem zamik, ki je bil izračunan s primerjavo



Slika 8: Odvisnost Pearsonove korelacije med napovedanim zadnjim tečajem in velikostjo intervala povprečja kombinirane ocene.

intervalov velikosti 312 do 332 ur in zamike velikosti od 10 do 35 ur. Na koncu se je izkazalo, da interval velikosti 322 ur in zamik velikosti 26 ur predstavljata vrednost z največjo korelacijo z napovedanim zadnjim tečajem, pri čemer je Pearsonov koeficient korelacije enak 0,6782.

6.3 Pearsonova korelacija med atributi

V tem razdelku so predstavljene končne Pearsonove korelacije med atributi, določenimi v prejšnjih razdelkih, ki izhajajo iz vrednosti delnic in tvitov. Atributi, ki se nanašajo na razpoloženje tvitov, ki so vsota kombinirane ocene, odstotek negativnih, nevtralnih in pozitivnih tvitov, so bili izračunani na podlagi intervala 322 ur in zamika 26 ur za vsak posamezen dan. Izračunan je bil Pearsonov koeficient korelacije z zadnjim tečajem. Kot je razvidno iz preglednice 2, ki prikazuje Pearsonove korelacije med napovedanim zadnjim tečajem in atributi, vidimo, da je atribut, ki predstavlja vsoto kombinirane ocene tvitov, najbolj koreliran z napovedanim zadnjim tečajem. V drugem položaju pa imamo odstotek pozitivnih tvitov s korelacijo enako 0,54. Po drugi strani pa ima odstotek negativnih tvitov negativno korelacijo z vrednostjo delnice. Odstotek spremenitve in nazadnje odstotek nevtralnih tvitov pa predstavljata atribute z najnižjo korelacijo z napovedanim zadnjim tečajem.

Tabela 2: Pearsonov koeficient korelacije med napovedanim zadnjim tečajem delnice in atributi, pri čemer upoštevan interval velikosti 322 ur in 26 ur zamika.

Atributi	Korelacija z napovedani zadnjim tečajem dneva t
Vsota kombinirane ocene dneva t	0,68
Odstotek pozitivnih tvitov dneva t	0,54
Odstotek negativnih tvitov dneva t	-0,49
Odstotek nevtralnih tvitov dneva t	0,09
Odstotek spremenitve dneva t-1	0,14

6.4 Analiza rezultatov

V trenutnem poglavju smo se osredotočili na določanje povezave med vrednostjo delnice in razpoloženje objav na omrežju Twitter. Na podlagi tehnik in strategij združevanja smo uspeli dokazati, da obstaja povezava med vrednostjo delnice TSLA in razpoloženje tvitov. To nam omogoča, da potrdimo drugo raziskovalno vprašanje, da je gibanje razpoloženja na omrežju Twitter povezano z gibanjem vrednosti delnice.

7 Napoved vrednosti delnice

Eden od načinov napovedovanja vrednosti delnice je napoved številčne vrednosti delnice naslednjega dneva, ki pravzaprav predstavlja glavni cilj tega poglavja. Poleg tega se bomo v tem poglavju osredotočili na raziskovalno vprašanje: Ali je napovedni model vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji samo na preteklih vrednostih delnice? Poglavje bo razdeljeno na dva glavna dela, ki predstavljata osnovno napoved in napoved z uporabo razpoloženja na Twitterju. Poleg tega pa bodo predstavljeni tudi koraki analize in razčlenitev časovne vrste zadnjega tečaja delnice, pri čemer se uvrščajo trend, sezonskost in cikli. Dosežki tega poglavja pa bodo predstavljeni v zadnjem razdelku, prikazali bomo vpliv razpoloženja na napoved vrednosti delnice.

7.1 Napovedovanje na osnovi podatkov o delnici

Analiza časovnih vrst je temeljni korak pri razumevanju in operiranju s časovnimi vrstami, ki nam omogočajo iskanje lastnosti časovne vrste za njeno napovedovanje. V naslednjih razdelkih bomo predstavili postopek razčlenitev časovne vrste vrednosti zadnjega tečaja delnice TSLA, pri čemer bomo analizirali trend, sezonskost in cikle.

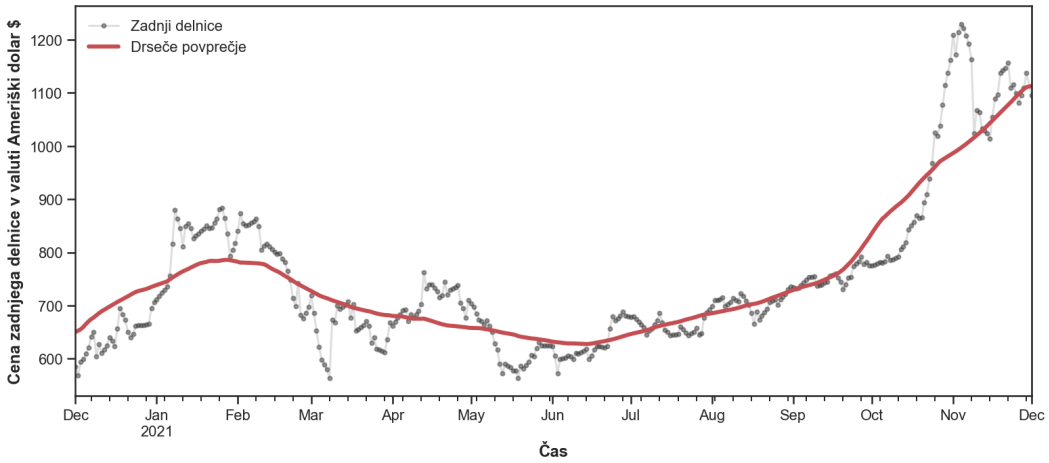
7.1.1 Analiza trenda

Trend časovne vrste predstavlja dolgoročno gibanje povprečja časovne vrste [28]. V tem razdelku, ki predstavlja prvi korak k dekompoziciji časovne vrste zadnjega tečaja delnice TSLA, bomo analizirali trend le-te.

Drseče povprečje

Eden od načinov za ugotavljanje trenda časovne vrste je uporaba drsečega povprečja [14]. Drseče poprečje temelji parametru, ki je velikost okna, pri čemer se le-ta premika vsako iteracijo eno vrednost naprej, na podlagi vrednosti, ki jih okno zajame, se izračuna povprečje le-teh [14]. Tako pridobljene vrednosti prikažejo novo časovno vrsto, ki označuje dolgoročno gibanje vrednosti delnice [14].
V našem primeru smo za okno vrednosti delnice TSLA izbrali 70 dni, ker najbolje

prikazuje dolgoročno gibanje vrednosti delnice, rezultat pa je prikazan na sliki 9. Kot je razvidno, drseče povprečje sledi kubičnem trendu, saj se vrednost delnice v prvem četrtletju premika najprej navzgor, nato pa v naslednjem delu upada, na koncu pa spet narašča in zajame največjo vrednost delnice v preučevanem obdobju.



Slika 9: Dršeče povprečje zadnjega tečaja delnice TSLA, pri čemer se je uporabilo okno velikosti 70 dni. Rdeča črta prikazuje le-tega, siva črta s točkami pa dejanski potek delnice v obdobju.

Napovedovanje trenda

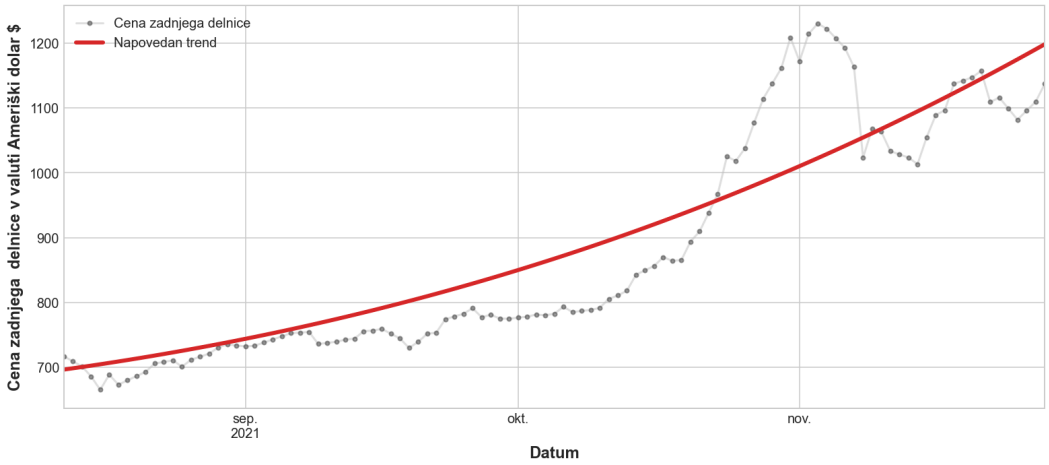
V prejšnjem razdelku smo ugotovili, da zadnji tečaj delnice sledi kubičnemu trendu. V trenutnem razdelku bomo predstavili napoved trenda delnice. Pri tem koraku smo razčlenili množico zadnjih tečajev delnice TSLA na 70% učno in 30% testno in napovedali trend z uporabo modela linearne regresije. Ena od glavnih značilnosti linearne regresije je, da lahko izražamo linearne polinome in tudi polinom n-te stopnje, ki nam omogočijo modeliranje trenda časovne vrste [14]. V našem primeru smo za ustvarjanje atributov, ki bi nam omogočili prilagajanje kubičnim trendom, uporabili funkcijo DeterministicProcess iz knjižnice statsmodels. Tovrstna funkcija nam omogoča dodajanje atributov za strojno učenje, glede na izbrani red funkcije trenda, ki nam zagotavlja modeliranje le-tega [14]. V našem primeru kubičnega trenda bo funkcija Detrministri-cProcess dodala še tri atribute, ki so naslednji:

- linearen trend,
- kvadratičen trend in
- kubičen trend

Ti atributi bodo omogočili algoritmu strojnega učenja, ki je v tem primeru linearna regresija, najti primerne utežni za modeliranja celotne časovne vrste [14]. Kot razred strojnega učenja smo izbrali zadnji tečaj delnice, ker je naš cilj napovedati trend le-tega. Napoved je bila ustvarjena z linearno regresijo, pri čemer smo ugotovil, da se zaradi kubičnega reda polinoma pojavi prekomerno prileganje. Da bi se temu izognili, smo posledično izbrali linearni model z uporabo Ridge regresije, pri čemer smo uporabili naslednje parametre:

- Reševalnik (ang. solver) = cholesky
- Alfa (ang. Alpha) = 100000
- Najvišja iteracija (ang. max iter) = 10000

Na sliki 10 je prikazana napoved trenda delnice TSLA.



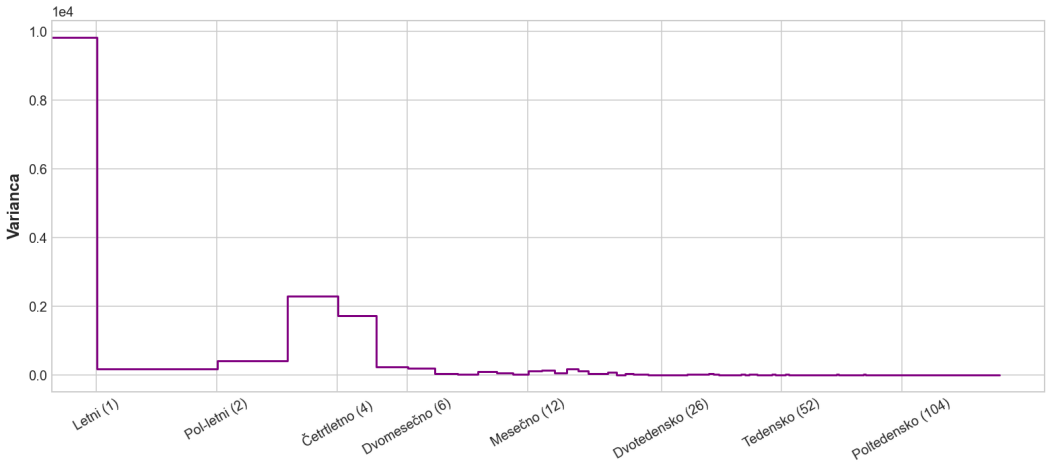
Slika 10: Napovedan trend zadnjega tečaja delnice TSLA z uporabo Ridge regresije, pri čemer rdeča črta prikazuje le-tega, siva črta s točkami pa dejanski potek delnice v obdobju.

7.1.2 Analiza sezonskosti

Sezonskost je izraz, ki se nanaša na ponavljajoča se gibanja, ki se ponavljajo v rednih časovnih presledkih [13]. Eden od načinov napovedovanja časovnih vrst se nanaša na analizo sezonskosti, torej na napovedovanje ponavljajočih se gibanj časovne vrste [13].

Fourierove vrste

Eden od načinov kreiranja atributov strojnega učenja za napoved sezonskosti je uporaba Fourierevih vrst. Fourierove vrste nam omogočajo aproksimacijo periodične časovne



Slika 11: Periodogram zadnjega tečaja delnice TSLA v obdobju med 1. decembrom 2020 in 1. decembrom 2021.

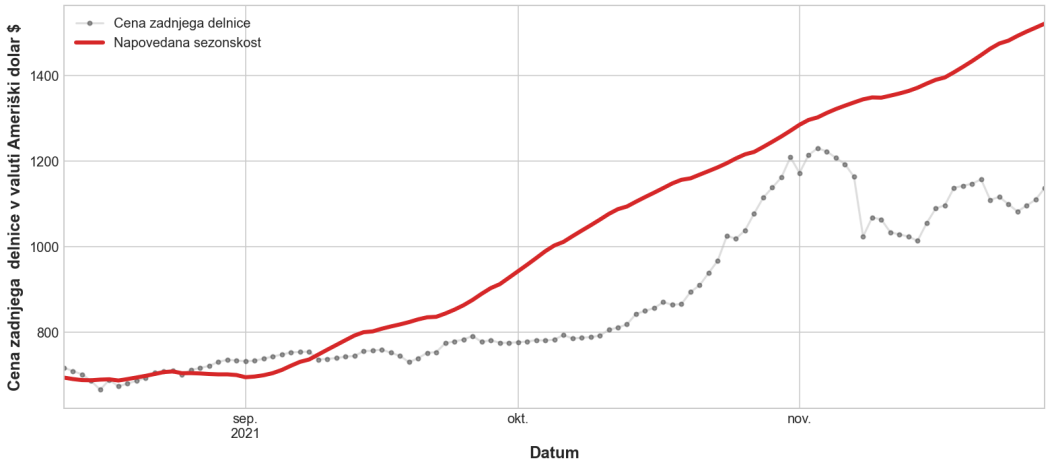
vrste z uporabo določenega števila parov sinusoidovih in kosinusodovih funkcij, večje bo število le-teh, večje bo tudi prileganje na vrednostih, ki jih model zavzame [13]. Eden od načinov za preverjanje prisotnosti sezonskosti je uporaba periodograma [13]. Periodogram je predstavitev frekvenc časovne vrste in njihovih moči. Vrednosti na y osi grafa je $(a^2 + b^2)/2$, kjer sta a in b koeficienta sinusa in kosinusa pri določeni frekvenci [13]. Pri analizi sezonskosti časovne vrste zadnjega tečaja delnice TSLA smo pridobili periodogram, ki je prikazan na sliki 11. Kot je razvidno, zadnji tečaj delnice TSLA kaže na četrtletno sezonskost, torej lahko modeliramo tovrstno sezonalnost na četrtletni ravni z uporabo 5 parov sinusnih in kosinusnih funkciji, ker nam omogoča, da najboljše aproksimiramo našo časovno vrsto. Za kreiranje atributov strojnega učenja smo uporabili funkcijo DeterministicProcess iz knjižnice statsmodels kot v primeru trenda. Tovrstna funkcija bo ustvarila atribute, ki bodo izbrano število parov sinusnih in kosinusnih funkciji, še dodatno ustvarila določeno število atributov vezanih na periodo časovne vrste [13]. Na podlagi teh vrednosti se algoritem strojnega učenja prilagodi časovni vrsti z izračunom uteži atributov, ki se nanašajo na komponento periode in sinusnih in kosinusnih komponent [13].

Napovedovanje sezonskosti in trenda

Na podlagi atributov časovnih vrst, ustvarjenih s funkcijo DeterministicProcess iz knjižnice statsmodels, smo razdelili dobljeno množico na 70% učno in 30% testno, pri čemer smo uporabili vrednost delnice kot razred. Kot model strojnega učenja smo ponovno uporabili regresijo Ridge, da bi se izognili prekomernemu prilagajanju, pri čemer smo uporabili naslednje parametre strojnega učenja:

- Reševalnik (ang. solver) = cholesky
- Alfa (ang. Alpha) = –10
- Najvišja iteracija (ang. max iter) = 10000

Rezultat napovedi sezonskosti skupaj s trendom je prikazan na sliki 12. Kot je razvidno, zadnji tečaj delnice nima močne sezonskosti, ki bi nam omogočila bolj natančno napoved vrednosti le-tega.



Slika 12: Prikaz napovedane sezonskosti in trenda zadnjega tečaja delnice TSLA, pri čemer rdeča črta prikazuje le-tega, siva črta s točkami pa dejanski potek delnice v obdobju.

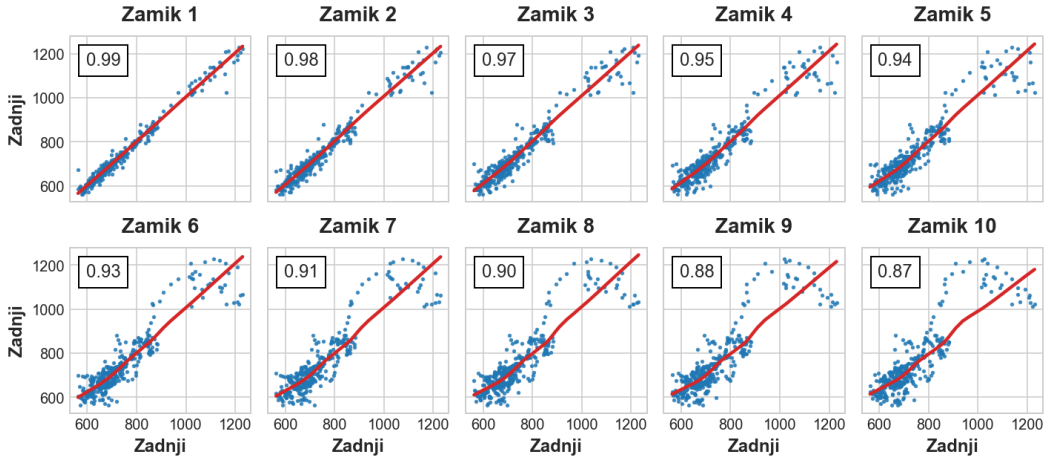
7.1.3 Analiza ciklov

Tretji korak, ki smo ga opravili pri napovedovanju vrednosti delnice, so bili cikli. Cikli se nanašajo na vrednosti časovne vrste, ki so odvisne od določenega števila prejšnjih vrednosti le-teh [12]. Ker so te vrednosti odvisne ena od druge, se lahko napoved naredi kot kombinacija teh vrednosti [12]. V tem razdelku bomo predstavili korake analize časovne vrste, pri čemer smo uporabili grafe zamika in globalen pristop in na podlagi le-teh ustvarili atribute strojnega učenja.

Grafi zamika

Eden od načinov preverjanja prisotnosti ciklov v dani časovni vrsti je izris grafov zamika (ang. lag plots). Graf zamika sestoji iz več raztrosnih grafov, pri čemer vsak posamezni raztrosni graf prikazuje na eni osi vrednosti časovne vrste do dneva x, na drugi osi pa zamaknjene vrednosti za določeno obdobje [12]. Na sliki 13 je prikazan graf zamika

zadnjega tečaja delnice TSLA, pri čemer smo uporabili zamik od 1 do 10 dni. V vsakem raztrosnem grafu je prikazana tudi številka avtokorelacije med časovno vrsto do dneva x in zapoznele časovne vrste. Kot je razvidno iz slike 13, so vrednosti zadnjega tečaja dne x, pri čemer je x poljuben dan, močno povezane z vrednostmi zadnjimi tečaji predhodnih vrednosti od x.



Slika 13: Graf zamika zadnjega tečaja delnice TSLA.

Napovedovanje vrednosti delnice

Po ugotovitvi, da so vrednosti delnice zadnjega tečaja med seboj povezane, smo poskusili napovedati zadnji tečaj delnice na podlagi prejšnjih zadnjih tečajev, pri čemer smo primerjali število upoštevanih dni za napoved z uporabo metrike RMSE. Bolj natančno – uporabili smo globalni pristop (ang.global approach), ki so ga opisali avtorji Parmezan A. in drugi v raziskavi [22]. Namreč, ta pristop uporablja vse vrednosti za ustvarjanje napovedovalnega modela, pri čemer uporabljamo drseče okno določene velikosti n. Na podlagi podzaporedji velikost n+1, ki jih drseče okno zajame, vsi razen zadnjega postanejo atributi, medtem pa zadnja vrednost postane razred strojnega učenja [22]. V končni fazi se dobljena množica razdeli na učno in testno in se ustvari napovedovalni model strojnega učenja [22]. Na sliki 3 je prikazan primer globalnega pristopa, pri čemer imamo vrednosti časovne vrste $z_1, z_2, ..., z_{10}$ in okno velikosti 3.

Nato je bil za napovedovanje vrednosti delnice uporabljen globalni pristop s primerjavo različnih velikosti oken oziroma števila predhodno izbranih zadnjih tečajev. Poleg tega smo primerjali tudi različne algoritme strojnega učenja za napovedovanje vrednosti zadnjega tečaja. Na sliki 14 so prikazane primerjave algoritmov strojnega učenja pri napovedovanju vrednosti delnice. Modeli strojnega učenja, ki so bili uporabljeni so

Tabela 3: Uporaba globalnega pristopa, pri čemer smo aplicirali 10 vrednosti časovnih vrst z velikostjo okna 3.

Atributi			Razred
z_1	z_2	z_3	z_4
z_2	z_3	z_4	z_5
z_3	z_4	z_5	z_6
z_4	z_5	z_6	z_7
z_5	z_6	z_7	z_8
z_6	z_7	z_8	z_9
z_7	z_8	z_9	z_{10}

naslednji:

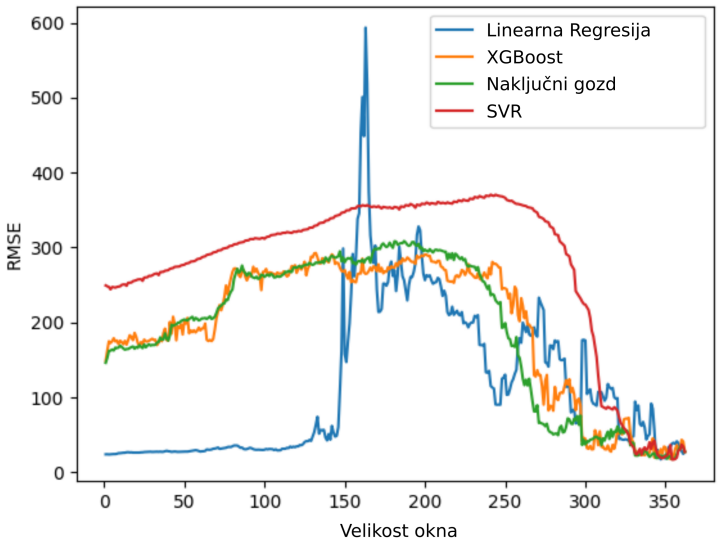
- linearna regresija,
- regresija XGBoost,
- regresija naključnih gozdov,
- regresija SVR.

Kot je razvidno, se je linearna regresija izkazala kot najboljša. Druga opazka pa je, da RMSE začne upadati po velikosti okna 250. To je verjetno poledica tega, da imamo po vsaki iteraciji eno vrstico manj in en atribut več. Torej po 364 dneh se napoved poveča na 364 atributov in en razred, tak imamo samo en primer.

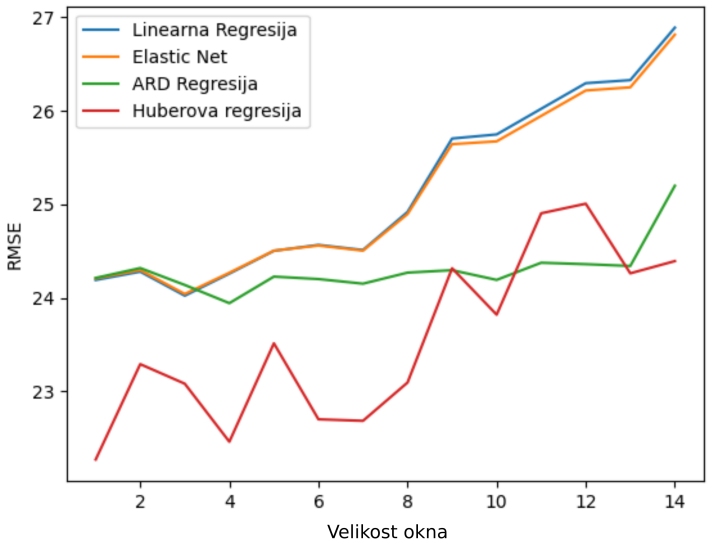
Zaradi boljšega doprinosa linearne regresije smo zato primerjali različne linearne modele strojnega učenja, rezultati pa so razvidni iz slike 15. Kot opazimo, se je Huberjeva regresija izkazal kot najboljši model linearne regresije pri napovedovanje vrednosti zadnjega tečaja delnice. Na podlagi tovrstnega algoritma strojnega učenja smo pridobili podatke napovedovanja, ki so razvidni v tabeli 4. V tabeli so predstavljene RMSE in R^2 napovedi vrednosti delnic, pri čemer je velikost predstavljenega okna od 1 do 5 dni. Poleg same primerjave velikosti okna so predstavljene primerjave atributov strojnega učenja. V prvi napovedi uporabili samo vrednosti zadnjega tečaja, v drugi napovedi pa poleg vrednosti zadnjega tečaja tudi odstotek spremenitve.

7.1.4 Hibridni modeli

Kot smo opazili, trend in sezonskost zadnjega tečaja delnice nista dovolj močna, da bi ju lahko uporabili za natančno napovedovanje vrednosti delnice, medtem ko so se



Slika 14: Primerjava algoritmov strojnega učenja pri napovedi vrednosti zadnjega tečaja delnice TSLA v odbobju med 1. decembrom 2020 in 1. decembrom 2021.



Slika 15: Primerjava linearnih algoritmov strojnega učenja pri napovedi vrednosti zadnjega tečaja delnice TSLA v obdobju med 1. decembrom 2020 in 1. decembrom 2021, omejena na okno velikosti 15.

cikli izkazali kot najboljši način za napovedovanje. Za nadaljnje izboljšanje napovedovanja smo se osredotočili na hibridne sisteme, ki nam omogočajo razčlenitev postopka napovedovanja na različne algoritme strojnega učenja [11]. V našem primeru smo prvi

Tabela 4: RMSE in R^2 napovedi zadnjega tečaja delnice TSLA z uporabo Huberjeve regresije, pri čemer smo primerjali velikosti okna in atribute strojnega učenja.

		Velikost okna.				
Atributi	Metrika	1	2	3	4	5
Zadnji tečaj.	RMSE	**22,268**	23,289	23,080	22,459	22,607
	R^2	**0,984**	0,982	0,982	0,983	0,983
Zadnji tečaj in odstotek spremenitve.	RMSE	23,327	23,252	**22,062**	22,431	22,484
	R^2	0,982	0,982	**0,984**	0,983	0,983

model uporabili pri napovedovanju z uporabo ciklov, drugi pa se uči na ostanku tega modela. Namreč, po napovedi prvega modela strojnega učenja in z uporabo naslednje enačbe:

$$y\_ostanek = y\_učna - y\_napoved$$

pri čemer je y_učni razred učne množice in je y_napoved napovedane vrednosti učne množice (x_učna), dobimo vrednosti y_ostanek [11]. y_ostanek predstavlja del učne množice, ki se ga model strojnega učenja ne uspe naučiti in postane učni razred drugega modela strojnega učenja, pri čemer uporabimo vrednost delnice kot atribut [11]. Končna napoved teh dveh modelov je sešteta v eno samo število, ki označuje napovedano vrednost delnice v določenem obdobju [11]. Ker je prvi model linearen, smo za drugi model izbrali enega bolj kompleksnih modelov, pri čemer smo primerjali SVR, naključni gozdovi in regresija XGBOOST. Na koncu se je SVR izkazal za najboljšega, saj je uspel še dodatno izboljšati napoved. Rezultati pa so prikazani v tabeli 5.

7.2 Napovedovanje z uporabo razpoloženja objav na omrežju Twitter

V tem razdelku predstavljamo rezultate pri napovedovanju delnice z uporabo razpoloženja objav na omrežju Twitter, pri čemer želimo ugotoviti, ali je napovedni model vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji samo na preteklih vrednostih delnice. Pri temu koraku se bomo osredotočili zgolj na cikle in hibridne modele, saj so se v prejšnjem razdelku 7.1 izkazali kot najboljši način za napovedovanje vrednosti delnice.

Tabela 5: RMSE in R^2 napovedi zadnjega tečaja delnice TSLA z uporabo hibridnega modela z uporabo Huberjeve regresije in SVR, pri čemer smo primerjali velikosti okna in atribute strojnega učenja.

		Velikost okna.				
Atributi	Metrika	1	2	3	4	5
Zadnji tečaj.	RMSE	**22,135**	22,102	22,000	22,447	22,493
	R^2	**0,984**	0,984	0,984	0,983	0,983
Zadnji tečaj in odstotek spremenitve.	RMSE	22,156	22,428	**21,790**	22,073	22,428
	R^2	0,984	0,983	**0,984**	0,984	0,983

7.2.1 Analiza ciklov

Kot smo opazili v prejšnjem razdelku, so se cikli izkazali kot najboljši način za napovedovanje vrednosti delnice v primerjavi z trendom in sezonskostjo. Poleg samih vrednosti delnice in atributov, ki izhajajo iz te vrednosti, bomo v tem razdelku analizirali in primerjali različne načine napovedovanja vrednosti delnice z uporabo atributov, ki izhajajo iz razpoloženja objava na omrežju Twitter. Namreč, pri tem koraku smo uporabili atribute izpeljane iz sentimentalne analize, ki so bili predstavljeni v poglavju 6. Atributi so bili potem aplicirani z uporabo algoritma strojnega učenja – Huberjeva regresija, pri čemer smo uporabili 70% za učno in 30% za testno množico. Napovedovanje vrednosti delnice je bila ocenjena z uporabo metrike RMSE in R^2, rezultati so prikazani v preglednici 6. V tovrstni preglednici so prikazane primerjave z velikostjo okna ter primerjava med atributi strojnega učenja. Primerjali smo pristope, ki temeljijo na sami vrednosti delnice s pristopom, ki temelji na vrednosti, ki izhajajo iz vrednosti delnice in razpoloženja. Atributi razpoloženja so sledeči:

- Odstotek negativnih, nevtralnih in pozitivnih tvitov.
- Kombinirana ocena tvitov.

Kot je razvidno iz preglednice 6, smo z uporabo sentimenta uspeli izboljšati napoved zadnjega tečaja delnice.

7.2.2 Hibriden model

Kot je razloženo v razdelku 7.1.4, po napovedi prvega modela strojnega učenja pridobimo določen ostanek, ki ga lahko uporabimo za izboljšanje končne napovedi. V našem

Tabela 6: RMSE in R^2 napovedi zadnjega tečaja delnice TSLA z uporabo Huberjeve regresije, pri čemer smo primerjali velikosti okna, atribute strojnega učenja, ki temeljijo na vrednosti delnice in razpoloženje objav Twitterja.

		Velikost okna.				
Atributi	Metrika	1	2	3	4	5
Zadnji tečaj.	RMSE	**22,268**	23,289	23,080	22,459	22,607
	R^2	**0,984**	0,982	0,982	0,983	0,983
Zadnji tečaj in atributi razpoloženja tvitov.	RMSE	22,264	22,447	**21,847**	21,960	22,058
	R^2	0,984	0,983	**0,984**	0,984	0,984
Zadnji tečaja in odstotek spremenitve in razpoloženje tvitov.	RMSE	22,275	22,270	**21,605**	21,916	22,108
	R^2	0,984	0,984	**0,985**	0,984	0,984

primeru smo uporabili hibridni model, sestavljen iz linearnega modela Huberjeve regresije in še SVR, ki je bil uporabljen na ostanku. Napoved je bila narejena z uporabo naslednjih atributov:

- Vsota kombinirane ocene.
- Odstotek negativnih, nevtralnih in pozitivnih tvitov.
- Odstotek spremenitve delnice.
- Zadnji tečaj delnice.

Da bi pridobili boljšo napoved, smo optimizirali parametre strojnega učenja Huberjeve regresije kot tudi SVR. Pri optimizaciji parametrov Huberjeve regresije smo uporabili sledečo konfiguracijo:

- Prileganje intercepcije (ang. fit interception) = True.
- Alfa (ang. Alpha) = 0,001.
- Epsilon = 1.
- Najvišja iteracija (ang. max iter) = 10000000.

Po drugi strani pa smo za algoritem SVR uporabili sledečo nastavitev parametrov:

- C = 0,0001.

- Epsilon = 0,000001.
- Gama (ang. Gamma) = 'scale'.

Končne vrednosti RMSE in R^2 so prikazane v tabeli 7.

Tabela 7: RMSE in R^2 napovedi zadnjega tečaja delnice TSLA z uporabo hibridnega modela z uporabo Huberjeve regresije in SVR in uporabe optimiziranih parametrov strojnega učenja. Uporabljeni atributi izhajajo iz vrednosti delnice in vrednosti razpoloženja objav omrežja Twitter.

		Velikost okna.				
Atributi	Metrika	1	2	3	4	5
Zadnji tečaj in odstotek spremenitv in razpoloženje tvitov.	RMSE	22,234	22,250	**21,253**	21,698	22,147
	R^2	0,984	0,984	**0,985**	0,984	0,984

7.3 Analiza rezultatov

V tem poglavju smo se osredotočili na napoved vrednosti delnice. Pri tem smo spoznali različne vrste napovedovanja, kot so trend, sezonskost in cikli, pri čemer smo opazili, da najnatančnejši način napovedovanja temelji na ciklih. Druga in tudi najpomembnejša opazka je, da je napovedni model vrednosti delnice, ki vključuje razpoloženje boljši od modela, ki temelji samo na preteklih vrednostih delnice, kar nam omogoča potrditev tretje hipoteze.

8 Napoved naraščanja in padanja vrednosti delnice

Eden od načinov napovedovanja delnice je napovedovanje rasti ali padca vrednosti delnice naslednjega dneva, pri čemer je ena od uporab le-tega kratkoročno vlaganje v delnice. Če je namreč napovedana vrednost delnice višja od trenutne, lahko delnico kupimo ali jo obdržimo, v nasprotnem primeru jo prodamo. V tem poglavju se bomo osredotočali na zadnjo raziskovalno vprašanje, ki zatrjuje, ali je napovedni model gibanja vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji samo na preteklih vrednostih delnice? V tem poglavju bomo predstavili potek preoblikovanja napovedovanja delnice v klasifikacijski problem in bomo predstavili rezultate napovedovanja naraščanja in padanja vrednosti delnice. Pri temu koraku bomo analizirali in primerjali napoved s samo uporabo vrednosti delnice TSLA in napoved uporab razpoloženja objav na omrežju Twitter. Nazadnje bomo predstavili hipotetični dobiček, ki naj bi bil dosežen z vlaganjem, če bi bile uporabljene napovedi modela.

8.1 Predobdelava podatkov

Pri klasifikaciji napovedujemo določeno vrednosti in zato je treba razred opredeliti vnaprej. Naš razred je v tem primeru binaren in sestavljen od ene vrednosti, ki bo kazala na rast vrednosti delnice in ene vrednosti, ki bo kazala na padanje vrednosti delnice v primerjavi s prejšnjim dnevom. V našem primeru smo se odločili za uporabo vrednosti 1, za naraščanje vrednosti delnice, v nasprotnem primeru pa vrednost –1. Zato je bil dodan nov stolpec v datoteki CSV, ki vsebuje te podatke, ki so bili izračunani na podlagi znanih vrednosti delnice. Stolpec pa je bil poimenovan gibanje vrednosti delnice.

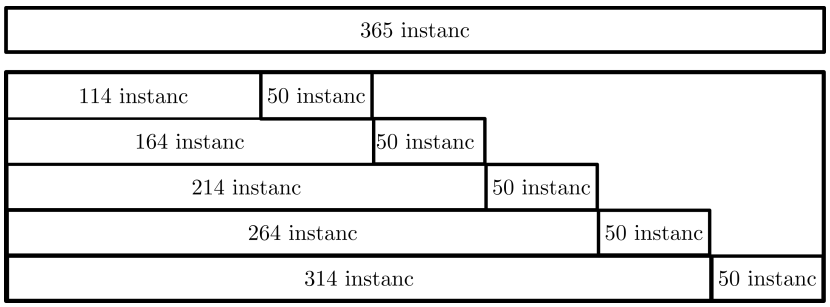
8.2 Napovedovanje gibanja vrednosti delnice

V tem razdelku bomo predstavili napoved gibanja vrednosti delnice, pri čemer bomo primerjali pristop, ki temelji na osnovi podatkov o delnici in napoved, ki temelji na razpoloženju objav na omrežju Twitter. Kot v prejšnjem poglavju 7 se bomo osredotočili

na napovedovanje zadnjega tečaja delnice za naslednji dan in natančneje napovedali gibanje, torej rast in padec delnice naslednjega dne.

8.2.1 Prečno preverjanje

Prvi korak pri napovedovanju vrednosti delnice, poleg podatkov, je določiti velikost učne in testne množice. V našem primeru razdelitev množice na 70% za učenje in 30% za testiranje privede do neuravnotežene množice in poleg tega testna množica postane preveč majhna. Zato smo uporabili pristop prečnega preverjanja, pri čemer smo razdelili množico 365 instanc v 5 učnih in 5 testnih množic, kot je prikazano na sliki 16. Z uporabo prečnega preverjanja v tem primeru bomo imeli 5 iteraciji, pri čemer bomo v vsaki iteraciji uporabili vedno večjo učno množico. V prvi iteraciji bomo uporabili učno množico veliko 114 instanc oziroma dni in 50 instanc za testno množico, v naslednji iteracijah pa bomo dodali prejšnjo testno množico učni množici in bomo uporabili naslednjih 50 instanc za testno množico. Z uporabo prečnega preverjanja opazimo, da je testna množica sestavljena iz 250 dni oziroma instanc.



Slika 16: Uporaba prečnega preverjanja na 365 primerov, pri čemer smo le-te razdelili na 5 učnih in 5 testnih množicah.

8.2.2 Napovedovanje na osnovi podatkov o delnici

Pri osnovnem koraku za napovedovanje gibanja vrednosti delnice smo se osredotočili na cikle časovne vrste in smo uporabili atribute, ki izhajajo le iz vrednosti delnice. Izbrali smo naslednje atribute, pri čemer napovedujemo gibanje vrednosti dneva t:

- Gibanje vrednosti delnice dneva t-1, t-2.
- Odstotek spremenitve delnice dneva t-1, t-2.

Algoritem strojnega učenja, ki je bil uporabljen pa je logistična regresija z uporabo prečnega preverjanja, opisan v razdelku 8.2.1.

Pri optimizaciji parametrov logistične regresije smo uporabili sledeče parametre:

- Prileganje intercepcije (ang. fit_interception) = true.
- Reševalnik (ang. solver) = 'saga'.
- Najvišja iteracija (ang. max_iter) = 10000.
- C = 90.

Pri napovedovanju pa smo izračunali aritmetično povprečje metrik uspešnosti vseh 5 napovedi testnih množic. Vrednosti metrik uspešnosti so pa prikazane v razpredelnici 8.

Tabela 8: Metrike uspešnosti pri napovedovanju zadnjega tečaja delnice TSLA z uporabo modela Logistične regresije in atribute, ki izhajajo iz zadnjega tečaja delnice TSLA.

Metrike uspešnosti	Vrednost
Točnost(ang. accuracy)	65,60%
Preciznost (ang. precision)	70,16%
Priklic (ang. recall)	67,09%
Mera F1 (ang. F1 score)	68,55%

8.2.3 Napovedovanje z uporabo razpoloženja objav na omrežju Twitter

V tem razdelku bomo predstavili napoved gibanja vrednosti zadnjega tečaja delnice z uporabo razpoloženja objav na omrežju Twitter. Tudi v tem primeru smo uporabili logistično regresijo in prečno preverjanje, predstavljeno v poglavju 8.2.1, pri čemer smo izbrali sledeče atribute pri napovedovanju dneva t:

- Gibanje vrednosti delnice dneva t-1 in t-2.
- Odstotek spremenitve delnice dneva t-1 in t-2.
- Kombinirana ocena sentimenta dneva t-1, t-2.
- Odstotek negativnih, nevtralnih in pozitivnih tvitov dneva t-1 in t-2.

Pri optimizaciji parametrov logistične regresije smo izbrali sledeče parametre:

- Prileganje intercepcije (ang. fit_interception) = true.
- Reševalnik (ang. solver) = 'saga'.

- Najvišja iteracija (ang. max_iter) = 10000.
- C = 45.

Rezultati napovedovanja so prikazani v preglednici 9, pri čemer so vrednosti metrike uspešnosti aritmetično povprečje napovedi testne množice.

Tabela 9: Metrike uspešnosti pri napovedovanju z uporabo modela logistične regresije in atributi, ki izhajajo iz delnice in razpoloženja objav na omrežju Twitter.

Metrike uspešnosti	Vrednost
Točnost (ang. accuracy)	67,99%
Preciznost (ang. precision)	71,49%
Priklic (ang. recall)	73,48%
Mera F1 (ang. F1 score)	72,25%

Kot je razvidno v preglednicah 8 in 9, uporaba razpoloženja objav na Twitterju prispeva k izboljšanju napovedi vrednosti delnice.

8.3 Simulacija zaslužka

Napovedi klasifikacijskega modela prejšnjega razdelka 8.2 se lahko uporabi pri dejanskemu trgovanju z delnicami pri napovedovanju prodaje ali nakupu/obdržanju delnice na določen datum. Če napovedovalni sistem napove –1 bomo naše delnice prodali, po drugi strani, če je napoved 1 bomo delnico kupili oziroma jo obdržali. V trenutnem razdelku bomo predstavili morebitni zaslužek na podlagi napovedovalnega modela.

8.3.1 Simulacija

Zaslužki so bili izračunani na podlagi testnih množic, ki je kot posledica prečnega preverjanja, razdeljena na 5 delov velikih 50 instanc oziroma dni. Pri simulaciji smo izbrali začetno naložbo v vrednosti 1000$. Na podlagi napovedi modela smo izračunali dobiček v vsakem delu, pri čemer se vrednost ohranja iz enega dela na drugega. Pri izračunu nismo upoštevali morebitne provizije pri prodaji, ali nakupu delnice. Rezultati primerjave napovedovalnih modelov so prikazani v preglednici 10, pri čemer smo primerjali različne pristope napovedovanja. Uporabili smo naslednje modele:

- Napovedovalni model, ki temelji na vrednosti delnice TSLA, ki je bil opisan v razdelku 8.2.2.

- Napovedovalni model, ki temelji na vrednosti delnice TSLA in razpoloženju objav na omrežju Twitter, ki je bil opisan v razdelku 8.2.3.
- Napovedovalni model, s 100-odstotno natančnostjo, ki vnaprej zazna, če bo vrednost delnice padla, ali narasla.

Kot je razvidno ima napovedovalni model, ki vsebuje poleg samih vrednosti delnice tudi razpoloženje objav na omrežju Twitter, poleg boljše točnosti in preciznosti tudi boljšo donosnost naložbe.

Tabela 10: Primerjava zaslužkov napovednih modelov z začetno naložbo 1000$.

Del	Napovedovalni model na podlagi prejšnjih vrednosti delnice in odstotka spremenitve delnice.	Napovedovalni model na podlagi prejšnjih vrednosti delnice, odstotka spremenitve delnice in vrednostnimi **razpoloženja** objav	Napovedovalni model s 100-odstotno natančnostjo
1	1036,78 $	1171,53 $	1548,93 $
2	1132,01 $	1389,81 $	2242,99 $
3	1260,14 $	1511,15 $	3075,73 $
4	1397,50 $	1679,70 $	3888,23 $
5	2066,15 $	2444,45 $	7810,94 $

Številne aplikacije za naložbe v delnice danes, kot so Robinhood[1], ne zaračunavajo provizije pri prodaji, ali nakupu delnice, zato se lahko delnico kupi, ali proda večkrat, ne da bi pri tem izgubili denar [26]. Po drugi strani pa na podlagi zakonodaje v Sloveniji stopnja davka znaša 27,5% na davčni osnovi od dobička od kapitala, če je prodaja izvršena v 5-ih letih od nakupa delnice [2]. Torej v našem primeru, pri uporabi modela napovedovanja, na podlagi prejšnjih vrednosti delnice in razpoloženja, dobiček znaša 1444,454$, pri čemer bi davek znašal 397,224$, dobiček pa bi znašal 1048$. To nam omogoča sklepanje, da bi tovrstni model napovedovanja v obdobju od 1. decembra 2020 do 1. decembra 2021 privedel do skupnega dobička 1048$, pri čemer predpostavimo,

[1]RobinHood – https://robinhood.com/us/en/

da je vsak dan v navedenem obdobju trgovalni in upoštevamo aproksimacijo vrednosti, ki je bila prikazana v razdelku 5.1.2.

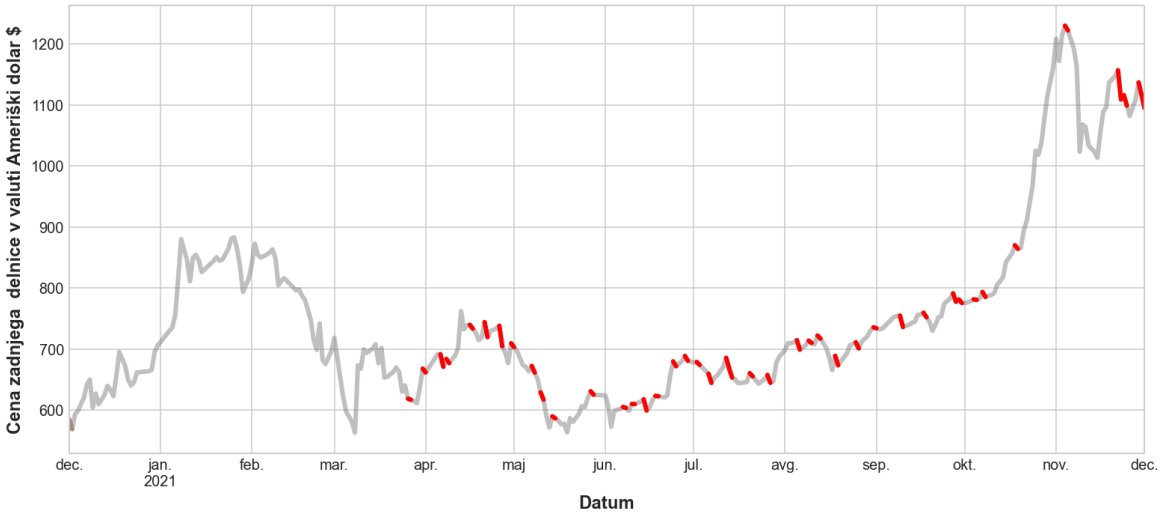
8.3.2 Grafični prikaz napak

Na podlagi modela, ki temelji na prejšnjih gibanj delnice in razpoloženja, ki je bil opisan v razdelku 8.2.3, bomo v tem razdelku izpostavili graf, ki prikazuje napake, ki jih je model naredi pri napovedovanju. Na sliki 17 je prikazan graf vrednosti delnice, ki je bil pobarvan na podlagi uspešne napovedi vrednosti delnice. Graf je pri dneh, kjer je bilo gibanje vrednosti delnice napovedana pravilno, označen z zeleno barvo, kjer je bila vrednost delnice napovedana nepravilno pa je na grafu označeno z rdečo barvo. Del, prikazan v sivi barvi pa predstavlja začetno učno množico, ki je bila uporabljena zgolj za učenje modela strojnega učenja.

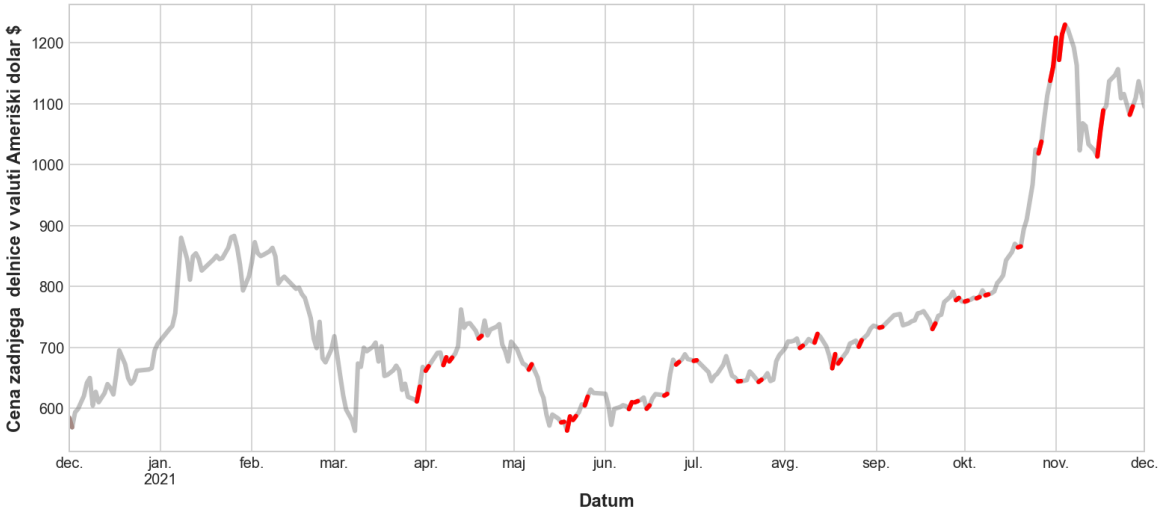
Slika 17: Grafični prikaz napak in pravilno napovedanih vrednosti zadnjega tečaja delnice TSLA, ki jih je model naredil pri napovedovanju gibanja vrednosti zadnjega tečaja TSLA. Siva barva prikazuje testno množico, ki je bila uporabljena le za učenje modela. Zelena in rdeča barva prikazujeta pravilno oziroma nepravilno napoved zadnjega tečaja delnice TSLA določenega dneva.

Na sliki 18 deli, označeni z rdečo barvo, prikazujejo dneve, ko je bila napovedana rast vrednosti delnice, v resnici pa je vrednost padla. Te napačne napovedi modela prikazujejo napovedi, ki vodijo do denarne izgube vlagatelja.

Na sliki 19 deli, označeni z rdečo barvo, prikazujejo dneve, ko je bil napovedan padec vrednosti delnice, v resnici pa je vrednost narasla. Te napačne napovedi modela prikazujejo napovedi, ki vodijo do izgub zaslužka vlagatelja.



Slika 18: Grafični prikaz napačnih rasti vrednosti delnice, ki jih je model naredil pri napovedovanju gibanja vrednosti zadnjega tečaja delnice TSLA. Tovrstne napake so prikazane v rdeči barvi, vse ostalo pa je prikazano v sivi barvi.



Slika 19: Grafični prikaz napačnih padcev vrednosti, ki jih je model naredil pri napovedovanju gibanja vrednosti delnice. Tovrstne napake so prikazane v rdeči barvi, vse ostalo pa je prikazano v sivi barvi.

8.4 Analiza rezultatov

Kot lahko opazimo na podlagi prejšnjih podpoglavji 8.2, je napovedovalni model gibanja vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji samo na preteklih vrednostih delnice, kar nam omogoča potrditev četrte hipoteze. Rezultati, z uporabo razpoloženja objav, kažejo na 67,99-odstotno točnost, ker je v skladu z dru-

gimi raziskavami na tem področju, kot je raziskava [21] avtorjev Sasank Pagolu V. in drugi, pri čemer so dosegli 69,01-odstotno točnost pri napovedovanju gibanja vrednosti delnice Microsoft z uporabo logistične regresije.

Poleg samega napovedovanja vrednosti delnice se je naš napovedovali model izkazal za učinkovitega in uporabnega pri vsakodnevnih naložbah v delnice v preučevanemu obdobju.

9 Zaključek

V zaključni nalogi smo se osredotočili na uporabo razpoloženja objav na omrežju Twitter za napovedovanje vrednosti delnice TSLA. Za dosego tega cilja smo uporabili množico 1.234.137 tvitov, ki so bili objavljeni na omrežju Twitter v obdobju med 1. decembrom 2019 in 1. decembrom 2021, ki so vsebovali ključne besede glede podjetja Tesla in delnice TSLA. Na tovrstnih tvitih je bila opravljena analiza razpoloženja z orodjem VADER za razumevanje razpoloženja uporabnikov. Poleg podatkov o tvitih smo za napovedovanje uporabili tudi podatke o delnicah, pri čemer smo se osredotočili na obdobje od 1. decembra 2020 do 1. decembra 2021. Polarnost razpoloženja tvitov je bila nato združena v enotne vrednosti za vsak posamezni dan, kar nam je omogočilo uporabo le-teh za nadaljnjo obdelavo. Naslednji korak je bil določiti in dokazati obstoj povezave med razpoloženjem objav Twitterja in vrednostmi delnice z uporabo Pearsonovega koeficienta korelacije. Pri tem koraku smo potrdili drugo raziskovalno vprašanje in namreč, da je gibanje razpoloženja na omrežju Twitter povezano z gibanjem vrednosti delnice TSLA. Po ugotovitvi obstoja povezave med delniškim trgom in vrednostjo delnice smo se posvetili napovedovanju delnice podjetja Tesla z uporabo pristopov strojnega učenja. Ta korak je bil razdeljen na dva dela, ki sta napovedovanje vrednosti in napovedovanje gibanja vrednosti delnice. V obeh primerih smo primerjali tudi pristop, ki temelji le na atributih, ki izhajajo iz delnice in pristop, ki temelji na atributih, ki izhajajo iz delnice in razpoloženja uporabnikov. Tudi v teh dveh primerih smo potrdili tretjo in četrto raziskovalno vprašanje in namreč, da je napovedni model vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji samo na preteklih vrednostih delnice, in da je napovedni model gibanja vrednosti delnice, ki vključuje razpoloženje, boljši od modela, ki temelji samo na preteklih vrednostih delnice.

Poleg samega napovedovanja vrednosti delnice smo testirali in postavili temelje za praktično uporabo napovedi gibanja delnice in ga uporabili pri trgovanju v preučevanem obdobju. Rezultati kažejo, da se tovrstni model lahko uporablja za ustvarjanje dobička v obravnavanem časovnem obdobju.

9.1 Omejitve projekta

V zaključni nalogi so opisali korake nastanka napovedovalnega modela do uporabe letega za napoved vrednosti zadnjega tečaja vrednosti delnice TSLA. Pri teh korakih smo se srečali z omejitvami, ki jih bomo opisali v tem razdelku. Prva med omejitvami je, da je napovedovalni model omejen le na eno delnico, ki je v tem primeru TSLA. Poleg tega smo za napovedovanje izbrali obdobje enega leta in namreč obdobje med 1. decembrom 2020 in 1. decembrom 2021. Ena od posledic tega je, da ima to obdobje določene značilnosti, ki lahko vplivajo na rezultate in metrike uspešnosti.

Dodatno pa smo se omejili le na združevanje vseh objav iz omrežja Twitter ne glede tipologije, ali avtorja objav. Obstajajo uporabniki, ki imajo ogromen vpliv na uporabnike kot tudi na delniški trg. Torej govorimo o osebah, ki imajo veliko pozornosti in tudi sledilcev. Po drugi strani imamo običajne uporabnike, ki tega nimajo. V našem primeru smo vse to upoštevali kot en sam tip objav in to združili s tehnikami, ki so bile prikazane v poglavju 6. Poleg tega smo upoštevali le objave v angleškem jeziku.

Omejitve so nastale tudi pri napovedovalnih modelih, ki so se osredotočali le na modele strojnega učenja, kot so linearna regresija, SVM in ostali, ki so bili uporabljeni v tej zaključni nalogi. Torej, niso bili uporabljeni statistični pristopi, ali pa novejše tehnologije, kot so nevronske mreže.

9.2 Možnosti za nadaljnje delo

V zaključni nalogi smo dokazali, da je mogoče z uporabo razpoloženja objav na omrežju Twitter izboljšati napoved modela strojnega učenja. Kljub temu ima to področje še veliko prostora za napredek. Ena od teh možnosti je spremeniti način združevanja razpoloženja objav na omrežju Twitter. Namreč, obstajajo tviti, ki so napisani s strani posameznikov, ki imajo veliki vpliv na javno mnenje, kot so tviti poznanih ekonomistov, ali glavnih direktorjev določnega podjetja. Po drugi strani pa imamo tvite, ki so bili zapisani s strani posameznikov, ki imajo zelo majhen vpliv na mnenje ostalih posameznikov. Eno od možnih nadaljnjih vprašanj je, kako in kdaj ti dve vrsti tvitov vplivata in sta povezani na vrednost in gibanje delnice?

Še dodatna razširitev trenutnega modela strojnega učenja za napovedovanje gibljaja vrednosti delnice predstavljenega v poglavju 8 je trgovanje v realnem času. V tem primeru bi algoritem strojnega učenja uporabil zadnje objavljene tvite in na podlagi razpoloženja teh objav ustvaril priporočila, ali neposredno trgoval z delnicami.

10 Literatura

[1] Družbeno omrežje twitter. Dosegljivo: `https://twitter.com/home?lang=en`, [Dostopano: 02. 04. 2022]. *(Citirano na strani 7.)*

[2] Finančna uprava slovenije. Dosegljivo: `https://www.fu.gov.si/zivljenjski_dogodki_prebivalci/odsvojil_sem_vrednostne_papirje_druge_deleze_ali_investicijske_kupone/`, [Dostopano: 02. 04. 2022]. *(Citirano na strani 46.)*

[3] Programski jezik python. Dosegljivo: `https://www.python.org/`, [Dostopano: 01. 04. 2022]. *(Citirano na strani 16.)*

[4] Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction. In *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015*, pages 1349–1354. Institute of Electrical and Electronics Engineers Inc., oct 2015. *(Citirano na straneh 5 in 21.)*

[5] Financial Sentiment Lexicon Analysis. *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018*, 2018-Janua:286–289, 2018. *(Citirano na strani 22.)*

[6] Li Bing, Keith C.C. Chan, and Carol Ou. Public sentiment analysis in twitter data for prediction of a company's stock price movements. In *Proceedings - 11th IEEE International Conference on E-Business Engineering, ICEBE 2014 - Including 10th Workshop on Service-Oriented Applications, Integration and Collaboration, SOAIC 2014 and 1st Workshop on E-Commerce Engineering, ECE 2014*, pages 232–239. IEEE, nov 2014. *(Citirano na straneh 5 in 19.)*

[7] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011. *(Citirano na straneh 1 in 4.)*

[8] Sushree Sasmita Dash, Subrat Kumar Nayak, and Debahuti Mishra. A Review on Machine Learning Algorithms. In *Smart Innovation, Systems and Technologies*, volume 153, pages 495–507. 2021. *(Citirano na strani 10.)*

[9] Epcresilience. Podatkovna zbirka dogodkov epcresilience. Dosegljivo: `https://www.epcresilience.com/insight/resources/disaster-database`, [Dostopano: 11. 04. 2022]. *(Citirano na strani 18.)*

[10] Henok Girma. A tutorial on support vector machine. Dosegljivo: `https://web.fs.uni-lj.si/lasin/wp-content/include-me/neural/doc/seminar6.pdf`, [Dostopano: 25. 04. 2022]. *(Citirano na straneh 12 in 13.)*

[11] Ryan Holbrook. Kraggle - analiza časovnih vrst - hybrid models. Dosegljivo: `https://www.kaggle.com/code/ryanholbrook/hybrid-models`, [Dostopano: 11. 04. 2022]. *(Citirano na straneh 37 in 38.)*

[12] Ryan Holbrook. Kraggle - time series as features - cycles. Dosegljivo: `https://www.kaggle.com/code/ryanholbrook/time-series-as-features`, [Dostopano: 11. 04. 2022]. *(Citirano na straneh 10 in 34.)*

[13] Ryan Holbrook. Kraggle - time series as features - seasonality. Dosegljivo: `https://www.kaggle.com/code/ryanholbrook/seasonality`, [Dostopano: 11. 04. 2022]. *(Citirano na straneh 10, 32 in 33.)*

[14] Ryan Holbrook. Kraggle - time series as features - trend. Dosegljivo: `https://www.kaggle.com/code/ryanholbrook/trend`, [Dostopano: 11. 04. 2022]. *(Citirano na straneh 9, 12, 30, 31 in 32.)*

[15] E Philip Howrey and E Philip Howrey. Are Stock Prices Random or Predictable ? Linked references are available on JSTOR for this article : Are Stock : Prices Random or Predictable ?*. 1(1):21–24, 1965. *(Citirano na strani 1.)*

[16] C.J. Hutto. Vader. Dosegljivo: `https://github.com/cjhutto/vaderSentiment`, [Dostopano: 09. 03. 2022]. *(Citirano na straneh 22 in 23.)*

[17] JustAnotherArchivist. snscrape. Dosegljivo: `https://github.com/JustAnotherArchivist/snscrape`, [Dostopano: 09. 03. 2022]. *(Citirano na strani 21.)*

[18] John Kordonis, Symeon Symeonidis, and Avi Arampatzis. Stock price forecasting via sentiment analysis on Twitter. *ACM International Conference Proceeding Series*, (November), 2016. *(Citirano na strani 20.)*

[19] Nasdaq. Trading hours for the nasdaq stock markets. Dosegljivo: `https://www.nasdaq.com/stock-market-trading-hours-for-nasdaq`, [Dostopano: 25. 06. 2022]. *(Citirano na strani 21.)*

[20] Maude Nguyen-The, Guillaume-Alexandre Bilodeau, and Jan Rockemann. Leveraging Sentiment Analysis Knowledge to Solve Emotion Detection Tasks. nov 2021. *(Citirano na strani 22.)*

[21] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. Sentiment analysis of Twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350. IEEE, oct 2016. *(Citirano na straneh 4, 20 in 49.)*

[22] Antonio Rafael Sabino Parmezan, Vinicius M.A. Souza, and Gustavo E.A.P.A. Batista. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information Sciences*, 484:302–337, may 2019. *(Citirano na strani 35.)*

[23] Daniel Pyeong Kang Kim, Jongwhee Lee, Jungwoo Lee, and Jeanne Suh. Elon Musk's Twitter and Its Correlation with Tesla's Stock Market. *International Journal of Data Science and Analysis*, 7(1):13, 2021. *(Citirano na straneh 5 in 6.)*

[24] Bo Qian and Khaled Rasheed. Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1):25–33, jan 2007. *(Citirano na strani 1.)*

[25] Aroussi Ran. yfinance. Dosegljivo: `https://github.com/ranaroussi/yfinance`, [Dostopano: 09. 03. 2022]. *(Citirano na strani 19.)*

[26] Robinhood. Trading fees on robinhood. Dosegljivo: `https://robinhood.com/us/en/support/articles/trading-fees-on-robinhood/`, [Dostopano: 25. 06. 2022]. *(Citirano na strani 46.)*

[27] Tchai Tavor and Sharon Teitler-Regev. The impact of disasters and terrorism on the stock market. *Jàmbá Journal of Disaster Risk Studies*, 11(1):1–8, jan 2019. *(Citirano na strani 17.)*

[28] Statistični urad Republike Slovenije. Desezoniranje časovnih vrst. Dosegljivo: `https://www.stat.si/dokument/486/Desezoniranje_casovnih_vrst.pdf`, [Dostopano: 13. 04. 2022]. *(Citirano na straneh 9, 10 in 30.)*

[29] Jake VanderPlas. In-depth: Support vector machines. Dosegljivo: `https://jakevdp.github.io/PythonDataScienceHandbook/05.07-support-vector-machines.html`, [Dostopano: 23. 05. 2022]. *(Citirano na straneh IX in 13.)*

[30] Wikipedia. Coefficient of determination. Dosegljivo: `https://en.wikipedia.org/wiki/Coefficient_of_determination`, [Dostopano: 25. 04. 2022]. *(Citirano na strani 15.)*

[31] Wikipedia. Confusion matrix. Dosegljivo: `https://en.wikipedia.org/wiki/Confusion_matrix`, [Dostopano: 25. 04. 2022]. *(Citirano na strani 14.)*

[32] Wikipedia. F-score. Dosegljivo: `https://en.wikipedia.org/wiki/F-score`, [Dostopano: 29. 04. 2022]. *(Citirano na strani 15.)*

[33] Wikipedia. Linear regression. Dosegljivo: `https://en.wikipedia.org/wiki/Linear_regression`, [Dostopano: 19. 05. 2022]. *(Citirano na strani 12.)*

[34] Wikipedia. Logistic regression. Dosegljivo: `https://en.wikipedia.org/wiki/Logistic_regression`, [Dostopano: 25. 04. 2022]. *(Citirano na strani 11.)*

[35] Wikipedia. Pearsonov koeficient korelacije. Dosegljivo: `https://sl.wikipedia.org/wiki/Pearsonov_koeficient_korelacije`, [Dostopano: 22. 04. 2022]. *(Citirano na strani 25.)*

[36] Wikipedia. Sentiment analysis. Dosegljivo: `https://en.wikipedia.org/wiki/Sentiment_analysis`, [Dostopano: 28. 05. 2022]. *(Citirano na strani 8.)*

[37] Wikipedia. Time series. Dosegljivo: `https://en.wikipedia.org/wiki/Time_series`, [Dostopano: 19. 05. 2022]. *(Citirano na strani 9.)*

[38] Fan Zhang and Lauren J. O'Donnell. Support vector regression. In *Machine Learning*, pages 123–140. Elsevier, 2020. *(Citirano na strani 13.)*